Field Experiments in Education in Developing Countries

Karthik Muralidharan[1]

Abstract: The study of education in developing countries has been transformed by the rapid increase in the feasibility and prevalence of field experiments over the past fifteen years. This paper comprises three main sections. First, it illustrates the very broad range of research questions regarding education in developing countries that have been addressed using field experiments, and summarizes the most important patterns of findings from this body of research. Second, it discusses some of the limitations of field experiments and strategies for mitigating them through better design. Third, it provides a practical toolkit on design, implementation, measurement and data collection, analysis, and interpretation of field experiments in education. The overall goal for this chapter is to serve as a reference for students, researchers, and practitioners by summarizing lessons learned, highlighting key open questions for future research, and providing guidance on how to design and implement high-quality field experiments in education in a way that maximizes what we learn from them.

JEL Codes: C93; H42; I21; I25; I28; O15

Keywords: Education, field experiments, development, experimental design, synthesis, toolkit

[1] UC San Diego, J-PAL, NBER, and BREAD. E-mail: kamurali@ucsd.edu:

I thank Abhijit Banerjee, Alejandro Ganimian, Asim Khwaja, and Abhijeet Singh for several useful comments and discussions. All views represented here are my own, and not of any organization I am affiliated with.

1. Introduction

Perhaps no field in development economics in the past decade has benefited as much from the use of experimental methods as the economics of education. The rapid growth in high-quality studies on education in developing countries (many of which use randomized experiments) is perhaps best highlighted by noting that there have been *several* systematic reviews of this evidence aiming to synthesize findings for research and policy in *just the past three years*. These include Muralidharan 2013 (focused on India), Glewwe et al. 2014 (focused on school inputs), Kremer et al. 2013, Krishnaratne et al. 2013, Conn 2014 (focused on sub-Saharan Africa), McEwan 2014, Ganimian and Murnane (2016), Evans and Popova (2015), and Glewwe and Muralidharan (2016). While these are not all restricted to experimental studies, they typically provide greater weight to evidence from randomized controlled trials (RCT's).

The reviews above are mostly written for policy audiences and aim to summarize the policy implications of the research on education in developing countries. In contrast, this chapter is mainly written for graduate students and young researchers aiming to conduct field experiments in education in developing countries. The chapter hopes to achieve two goals. The first is to illustrate the broad range of studies that have been conducted in this area by providing a curated summary of the main insights from the research over the past fifteen years. The second (and more important) goal is to provide a practical toolkit on design, implementation, measurement and data collection, analysis, and interpretation of field experiments in education. The chapter aims to serve as a reference for students, researchers, and practitioners to guide the design and implementation of high-quality field experiments in education in a way that maximizes what we learn from them.

The chapter is organized as follows. Section 2 provides a conceptual overview of the core research questions in education in developing countries, and the role of field experiments in answering them. Section 3 highlights some of the key research questions in education in developing countries that have been addressed by field experiments, and aims to synthesize the main insights that have been obtained from this research in the past fifteen years. Section 4 discusses some of the important limitations of field experiments and ways in which they may be mitigated. Section 5 presents an extensive set of guidelines for students and practitioners on the effective design, implementation, data collection, analysis, and interpretation of field experiments. Section 6 concludes and discusses areas for future research.

2. Field Experiments in Education - a short overview

2.1. Background

Education and human capital are widely considered to be essential inputs for both aggregate growth and development (Lucas 1990; Barro 1991; Mankiw, Romer, and Weil 1992), as well as for enhancing the capabilities and freedoms of individuals, and thereby enabling them to contribute to and participate in the process of economic development (Sen 1993). Thus, improving education outcomes in developing countries has been an important policy priority for both national governments as well as for the international development and donor community. The importance of education in the development agenda is perhaps best reflected by the fact that two of the eight United Nations Millennium Development Goals (MDG's) pertained to education (achieving universal primary education, and achieving gender parity at all levels of education - both by 2015). The post-2015 Sustainable Development Goals (SDG's) continue to prioritize inclusive and equitable quality education for all (goal 4).

Further, education (especially school education) in most countries is typically provided publicly by the government and financed by taxpayer contributions. It is beyond the scope of this chapter to analyze *why* this is the case, but there are at least three reasons for the preponderance of publicly financed and provided education. First, there is considerable evidence to suggest that a variety of supply and demand-side constraints may prevent optimal education investments by households and optimal provision by markets, and that outcomes can be improved by a social planner (see Glewwe and Muralidharan 2016 for a review). Second, it is widely believed that education generates positive spillovers beyond the returns that accrue to individuals, which would also suggest an active role for governments in education financing and production.[2] Finally, an important non-economic reason for publicly-provided education may be states' desire to control curriculum and the content of education, which affect the formation of preferences and civic identity (Kremer and Sarychev 2008).

Thus, financing and producing education is an important policy priority for most countries, and public spending on education is typically one of the three largest components of government budgets (the other two being defense and healthcare). However, while this is true for most countries, developing countries face especially acute challenges in achieving universal quality education. They have lower levels of school enrollment and completion, much poorer

[2] Models with complementarity in worker human capital in production (such as Kremer 1993) predict spillovers. Lucas (1988) argues that human capital spillovers may be large enough to explain most of the long-run differences in per-capita income between high and low-income countries. Moretti (2004) provides evidence of such spillovers in the context of US manufacturing workers and plants. One direct channel of spillovers for which there is evidence in a developing country context is that education promotes technology adoption, and that non-adopters (who may be less educated) learn from adopters, which is a positive spillover from education that would not have been accounted for in the decision-making of individually optimizing agents (Foster and Rosenzweig 1995).

learning outcomes, and also have fewer public resources to spend on education. Thus, spending scarce public funds effectively is especially important in developing countries, where the opportunity cost of poor spending is higher.[3] As a result, a major area of focus for research on education in developing countries has been to understand the effectiveness (both absolute and relative) of various policy options to improve education outcomes.

2.2 The main research questions

Most experimental research on education in developing countries in the past decade has focused on two main policy questions. First - how should we increase school enrollment and attendance, and second - how should we improve learning outcomes? The two are closely related because increased enrollment and attendance are likely to be necessary pre-conditions for improving learning outcomes. Nevertheless, it is useful to think about the two problems distinctly because the school attendance decision is typically made by parents, whereas the extent to which increased school participation translates into improved learning outcomes is more likely to be affected by school-level factors.

On school participation and attendance, a simple model of optimizing households in the tradition of Becker (1962) and Ben-Porath (1967) yields the result that households will only invest in an additional year of education for their child if the present discounted value of the expected increase in benefits exceeds the costs of doing so. Thus, policies that seek to improve school participation typically aim to increase the immediate benefits to households of sending their children to school or to reduce the costs of doing so. The magnitude of the impact of these policies will in turn depend on the distribution of the household-child specific unobservables that determine whether a given child enrolls in or attends school, and the extent to which the policy helps make it more attractive to do so.

On quality of learning, a standard education production function with certain additional assumptions (see Todd and Wolpin (2003) for a detailed exposition of these assumptions), allows the lagged test score to be treated as a sufficient statistic for representing prior inputs into learning, and for the use of a value-added model to study the impact of changing contemporaneous inputs into education on test scores. Specifically, the typical value-added model takes the form:

$$T_{i,t} = \gamma T_{i,t-1} + \boldsymbol{\beta X}_{i,t} + \varepsilon_{i,t} \tag{1}$$

[3] In principle, governments in developing countries should be able to borrow to undertake any investment where the social rate of return is greater than the cost of borrowing. In practice, financial markets typically constrain the extent of government borrowing which places a hard budget constraint on public expenditure.

where $T_{i,t}$ represents test scores of child *i* at time t, $T_{i,t-1}$ represents the lagged test score, and $\boldsymbol{X}_{i,t}$ represents a full vector of contemporaneous home ($\boldsymbol{H}_{i,t}$) and school ($\boldsymbol{S}_{i,t}$) inputs. While the production function above is linear in $\boldsymbol{X}_{i,t}$ and is typically estimated this way, the specification does not have to be as restrictive, because $\boldsymbol{X}_{i,t}$ can include non-linear terms in individual inputs, and also include interaction terms between specific sets of inputs.

Given budget constraints in public education, and the almost unlimited set of ideas for inputs and interventions that may improve education outcomes, an optimal policy approach to allocating scarce resources across the set of potential inputs would be to estimate the marginal return to providing a specific input and to compare it with the marginal cost of doing so (since these inputs are typically provided publicly) and to prioritize investments in diminishing order of the estimated return per dollar spent. Further gains in effectiveness of education spending may be obtained by pivoting existing expenditure away from less to more cost-effective expenditure items.

Since cost data is relatively easier to obtain,[4] the main practical challenge is one of estimating the marginal returns to different inputs. The economics of education literature has correspondingly devoted a lot of attention to doing this and produced hundreds of papers across several developing countries trying to estimate these returns for various inputs (see Glewwe et al. 2014, and Glewwe and Muralidharan 2016 for a review).

2.3 The value of experiments in education and reasons for their growth

The main challenge for non-experimental studies (that use observational data) is the concern that variation in the specific input being studied $(X_{i,t})$ is correlated with the unobserved error term ($\varepsilon_{i,t}$), yielding biased estimates of β. In practice, this is quite likely to be true. For instance, communities and parents that care more about education are likely to be able to successfully lobby for more school inputs, and are also likely to provide unmeasured inputs into their children's education, which would lead to an upward bias on β estimated in cross-sectional data. In other cases, governments may target inputs to disadvantaged areas to improve equity in which case areas with increases in $X_{i,t}$ may be negatively correlated with $\varepsilon_{i,t}$, yielding downward biased estimates of β.

Thus, the value of experimental evaluations in this setting is quite clear since random assignment of the input (intervention) of interest solves this identification problem by ensuring that variation in $X_{i,t}$ is orthogonal to variation in $\varepsilon_{i,t}$, thereby yielding unbiased estimates of β

[4] In practice, even obtaining cost estimates of specific interventions is non-trivial (especially when it involves aggregating expenditure across multiple levels of government), but in principle they can be reconstructed from government budget documents.

(with some caveats as noted in section 4).[5] The importance of accounting for omitted variable bias in the evaluation of education interventions is starkly illustrated by Glewwe et al. (2004) who compare retrospective and prospective studies of the impact of classroom flipcharts on learning outcomes. When they use observational data, they find that flipcharts in classrooms appear to raise student test scores by 0.2σ. However, when they conduct a randomized controlled trial of flipcharts in classrooms, they find no impact on test scores at all, suggesting that the non-experimental estimates were significantly biased upwards (even after controlling for observable factors). These results underscore the value of field experiments for program evaluation in developing countries and Glewwe et al. (2004) can be considered analogous to LaLonde (1986) in the US program evaluation literature, which showed that non-experimental methods were not able to replicate the estimates from experimental evaluations of the impact of job training programs.

While field experiments have improved causal inference in most topics in applied micro-economics, they have been particularly prevalent in the economics of education (especially in developing countries) in recent years. There are several reasons for this. First, many interventions in education are "modular" and therefore feasible to randomize at the student, classroom, or school level. Second, the outcome variables are quite well defined and there is considerable agreement among economists on the key outcomes that programs should aim to improve (enrollment, attendance, and test scores[6]). Third, the large number of non-profit organizations that work on education has made it feasible for researchers to find implementation partners who can design and deploy the interventions being studied. Fourth, since non-government implementation partners typically cannot (and are not expected to) work "everywhere", it is politically and practically feasible for them to use a lottery to determine where their programs will first be rolled out since this ensure fairness in program access in addition to enabling experimental evaluations of impact. Finally, evidence from regularly conducted nationwide surveys like ASER in India and Uwezo in East Africa shows that despite large increases in school enrollment in developing countries, learning outcomes in these settings are very low (with a large fraction of students not being functionally literate or numerate at the end of primary school). The wide dissemination of these results has increased the demand from policy makers and funders of education programs for evidence on the impact of the programs they are funding, and for cost-effective ways of improving learning.

This confluence of factors has led to several high-quality experimental studies in education in developing countries that have both contributed evidence on the effectiveness of specific

---

[5] See the companion chapter by Athey and Imbens in this volume for more details (Imbens and Athey 2016).

[6] While test scores are not the ultimate outcomes that a social planner cares about, evidence using long-term longitudinal data find that interventions that raise test scores in school (such as having a better teacher in primary and middle school) also lead to better long-term labor market outcomes (Chetty et al. 2011).

programs and also promoted a deeper understanding of the barriers to improving education outcomes in these settings. These include studies on interventions to improve parent and student demand for education, school and household inputs, classroom organization and pedagogy, and school governance and teacher effort (with some interventions combining features across this broad classification). When put together (with some caveats), this body of research also enables comparison of marginal costs and benefits across different kinds of education spending and can guide policy priorities over allocation of limited resources.

2.4 An Alternative Framing of the Questions of Interest

The use of experiments above has been motivated by wanting to measure the "impacts" of interventions and to estimate their cost effectiveness at improving school participation and learning outcomes. Yet, an alternate way of framing the question of interest is to ask: "What are the determinants of school participation and learning outcomes?" This approach focuses on understanding household decision making regarding human capital investments as a function of household beliefs and preferences, and constraints including production function, budget, credit, and information constraints. Randomized evaluations of interventions are then interpreted less through the lens of their "impact" on outcomes of policy interest, and more through the lens of providing exogenous variation in the constraints above, which enables the researcher to better understand the determinants of school participation and learning outcomes (see Attanasio 2015 for an illustration of such an approach).

One way of thinking about the difference in these two ways of framing the question is that the approach in this section is more that of a "scientist" trying to understand the world, whereas the approach in section 2.3 is more that of an "engineer" trying to improve outcomes and solve problems (see Mankiw 2006 for a discussion of a similar distinction in approaches to macroeconomics). The synthesis of evidence in section 3 follows the approach in section 2.3 because most of the experimental research in education in recent years has been motivated by policy questions of how best to improve education outcomes in specific settings (this is also the style taken by the other systematic reviews referenced in the introduction). At the same time, there are important complementarities between the two approaches, and I argue in section 5 that studies that bridge this divide effectively will typically generate more generalizable insights, and mitigate against some of the limitations of experiments discussed in section 4.

3 Selected overview of field experiments in education in developing countries

As discussed earlier, this section therefore does not aim to provide a comprehensive review of field experiments in education in developing countries (see Glewwe and Muralidharan 2016 for such a treatment), but rather aims to illustrate the breadth of topics studied in this literature, and the broad patterns in the results to date. To organize the discussion below, I classify the range of interventions discussed into four broad categories: (1) those that are intended to increase the demand for schooling by students and their parents; (2) those that provide standard educational inputs through schools; (3) those that are related to changes in pedagogy; and finally (4) those that are related to the governance of schools, and to education systems more broadly. In the discussion below, I use the terms “experiment” and “RCT” (randomized controlled trial) interchangeably.

3.1 Demand-side interventions

The logic of demand-side interventions to improve education outcomes is that households may sub-optimally demand too little education for their children. Reasons include not accounting for spillovers from the education of their children to the overall economy, discounting the future at a higher rate than a social planner, having incorrect beliefs about the returns to education, and being credit constrained and unable to borrow for education even though investing in education would have a positive return. Thus, demand-side interventions aim to correct some of these sources of sub-optimal education investments.

Perhaps the most widely studied demand-side intervention using RCT's has been the idea of "conditional cash transfers (CCT)" (with eligibility often targeted to poorer households) whereby households receive a regular cash supplement if their children are enrolled in school and maintain a minimum attendance rate. While CCT programs aim to provide income support to the poor more generally (and not just increase demand for education), they have been found to have significant positive impacts on school enrollment and attendance across most settings where they have been evaluated using an RCT (see Fiszbein and Schady 2009 for a review of several of the early studies).

RCT's have also been used to study whether modifying the design of cash transfer programs can improve (or expand) the impacts of these initiatives. For instance, Baird et al. (2011) study the importance of conditioning cash transfers on school enrollment by comparing a standard CCT to an unconditional cash transfers (UCT) in Malawi and find that CCTs increase school enrollment by a larger amount than UCTs, but UCTs do better at protecting vulnerable girls by providing them with income even if they drop out of school. Similarly, Benhassine et al. (2013) find that labeling a UCT as being for education (in Morocco) achieved significant gains in school participation, and that adding conditionality did not yield any additional gains in schooling (though it added additional costs of monitoring and enforcing the conditionality). Finally, Barrera-Osorio et al. (2011) find that postponing part of the monthly transfer of a CCT to the time when school reenrollment has to take place (which is when fees need to be paid) significantly raised enrollment relative to a standard CCT in Colombia.

In addition to studies evaluating the impact of CCT's on school participation, the randomized roll-out of CCT programs across individuals and communities (most notably PROGRESSA-*Oportunidades-Prospera* in Mexico) has also enabled well-identified studies on important determinants of education participation including peer effects (Bobonis and Finan 2009, Lalive and Cattaneo 2009), consumption smoothing across households within communities (Angelucci and De Giorgi 2009), and the role of income controlled by women/mothers on children's consumption and education (Bobonis 2009). Finally, Todd and Wolpin (2006), and Attanasio et al. (2012) combine a structural model with PROGRESSA's experimental variation to generate predictions on schooling impact of the program under different values and design of the CCT program. Overall, CCTs have been one of the most highly-researched and deployed policy options to improve demand for schooling in developing countries, and have been a poster child for the value of carefully randomized program rollouts in generating high-quality evidence on both program impact, as well as on deeper determinants of household schooling investments.

A second prominent category of demand-side interventions studied experimentally relate to the provision of better information about education to students and parents. Since education decisions are taken on the basis of *perceived* as opposed to actual returns (Majumdar 1985), households may make sub-optimal decisions on education investments if they misperceive these returns. Jensen (2010) uses household survey data in the Dominican Republic to show that the perceived returns to high-school are much lower than the actual returns, and shows experimentally that simply providing students in randomly selected schools better information on the higher measured returns to secondary schooling led to a significant increase in the years of school completed. In a variant of this experiment, Jensen (2012) shows that providing randomly selected villages in northern India with information on the job opportunities available to educated young women and helping them access these opportunities (without changing the qualifications or standards for being hired) led to a significant increase in female education, and to delays in marriage and fertility.

Finally, Loyalka et al. (2013) conduct an experimental evaluation of the impact of providing information on returns to education, and career counseling services (in separate non-overlapping treatments) to junior high school students in China. They find that the information treatment had no impact on high-school participation, and also find that the career counseling treatment actually increased school dropouts and reduced test scores. The authors attribute this surprising negative result to the fact the wages of unskilled workers were rapidly rising in China in this period, and the possibility that the counseling services may have made academically weaker students decide that the academic requirements of higher education were too onerous and that it may make more sense for them to drop out and join the labor force.[7] The difference in the results of similar interventions across country contexts highlights the importance of caution in comparing results across contexts. It is also important to recognize that the impact of information will likely vary as a function of the prior beliefs and the extent and direction in which the information moved these beliefs.

RCT's have also been used to study the impact of providing information to students and parents about students' learning levels, with the idea being that parental and student investment in education is a function of their beliefs about the students' academic ability and that a misperception of true ability may lead to sub-optimal education investment decisions. Prominent examples include Dizon-Ross (2016) in Malawi and Bobba and Frisancho (2016) in Mexico. Both studies find evidence of mismatch as well as evidence of behavioral responses to the provision of information that is consistent with households updating their decisions in response to changes in their beliefs.

RCTs have also been used to study the impact of providing information on school quality in competitive education markets with multiple providers (both public and private). Andrabi et al. (2015) use a large-scale RCT across 112 villages in Pakistan to study the impact of providing parents with detailed student and school-level report-cards with information on test scores. They find that the intervention increased mean test scores by 0.11σ, reduced mean private schools by 17%, and also increased primary school enrollment by 4.5%. They also find that the mechanism for these results was an improvement in quality among the lower-quality schools and a reduction in price among the higher-quality schools, which is consistent with the predictions of models of optimal endogenous pricing and quality choice by providers in settings of asymmetric information (and how these should change in response to provision of better market-level information on quality)[8]. This study highlights the capacity of RCT's to yield *market level* insights on how the provision of information can affect parental demand and increase competitive pressure on schools and change outcomes.

A final category of demand-side interventions with promising experimental evidence on positive impacts is the provision of student-level incentives for better academic performance. Two prominent studies include Kremer, Miguel, and Thornton (2009), and Blimpo (2014). The first study conducts an RCT of a girls' merit scholarship in Kenya and finds significant positive effects on girls' test scores and also finds that teacher absence is reduced in treatment schools. Similarly, Blimpo (2014) conducts an RCT of three different types of student incentives in Benin (one based on individual incentives and two based on team incentives), and finds that all three

[7] Note that this result is similar to that seen non-experimentally by Atkin (2014) who shows that Mexican high-school students were more likely to drop out from school during a period of increasing demand for unskilled labor.
[8] They also verify that parents' knowledge of school quality did improve as a result of the intervention.

variants had a significant positive impact on the high-school exit exam test scores. Finally, Hirshleifer (2015) presents experimental evidence of the relative impact of rewarding students on the basis of education inputs (measured by their performance on practice exercises) versus rewarding them on the basis of education outputs (measured by their test scores), and finds that students with input rewards significantly outperform those with output rewards.

At the same time, it is important to note that demand-side interventions are not uncontroversial. For instance, the provision of information may make recipients of the information worse off if it is not correct. Specifically, one concern with the approach in Jensen (2010) is that the information provided on Mincerian returns to education may be incorrect on average (due to omitted variable bias) and also not be correct for the marginal student, since the returns to education for the marginal student induced to stay in school by the intervention are likely different from those to the average student (Carneiro, Heckman, and Vytlacil 2011).[9] Similarly, opponents of student incentives express concern that rewarding students for test score gains may crowd out students' intrinsic motivation for learning and acquiring knowledge for its own sake.

More generally, demand-side interventions tend to be paternalistic in nature since they typically assume that households are making sub-optimal education choices and need to be induced to demand more education. On the one hand, there is considerable evidence that there are important demand-side market failures that may lead to sub-optimal investments in education by parents and children. For instance, the large positive impacts of relatively small student prizes and incentives (which are especially small relative to the lifetime returns to completing schooling) suggest that students may underestimate the returns to education (or discount the future at a significantly higher rate than a social planner would). On the other hand, it is also possible that seemingly sub-optimal choices may be optimal in a local context and that a well-intentioned demand-side intervention may make people worse off. Thus, it is good practice for designers of demand-side interventions to check that they are responding to demonstrated evidence of sub-optimal choices in the specific context where the intervention is being considered, as opposed to simply assuming that this is true.

Overall, the evidence from experimental evaluations on the impact of demand-side interventions suggests that well-designed demand-side interventions are likely to be a promising avenue to explore for improving education outcomes in developing countries. Some interventions like the provision of better information are inexpensive and easy to scale up. Others like CCT's are expensive and much less cost-effective in terms of the additional years of

[9] Note that the approach in Jensen (2012) is less susceptible to this concern because the standards for hiring candidates did not change (and hence no potentially misleading information was provided) but new information was provided by the recruiting firms

schooling obtained per dollar spent (Dhaliwal et al. 2013).[10] Key open questions include experimenting with alternative designs of demand-side interventions to better match the intervention to the source of inefficiency and doing so in the most cost-effective manner.

3.2 School and Student Inputs

The vast majority of public education spending is devoted to school inputs including infrastructure, teacher salaries, and student inputs including textbooks and other learning materials. A considerable amount of education research has aimed to study the impacts of these inputs on school participation (enrollment and attendance) as well as learning outcomes (see Glewwe and Muralidharan 2016 for a review of this research).

However, there are relatively few RCT's of school construction and infrastructure construction since it is not easy to randomize construction of durable assets. Burde and Linden (2013) experimentally vary the creation of village-based schools in rural Afghanistan and find large effects of having a school in the village on school participation and test scores – especially for girls. They also use the exogenous variation in distance to the nearest school induced by their experiment and estimate that the distance-elasticity of school attendance is quite large – again, especially for girls.[11] Thus, school construction – especially building a school in villages that have no school – is likely to improve enrollment (and potentially test scores as well, relative to the counterfactual of not attending school). On the other hand, it is also true that the construction of schools in every village (or even hamlets within large villages) has led to the creation of many sub-scale schools with (especially in India). Thus, a key open question with regards to school access is that of the optimal trade-off between access and scale, and whether it is better to have fewer, larger schools that are better equipped and managed; with transport subsidies to ensure access to students living outside a walking distance to the school.[12]

Unlike infrastructure, there have been several experimental evaluations of the impact of providing schools with books and materials (or grants to be used for books and materials), and a consistent finding across studies has been that the provision of these materials *on their own* does not lead to significant improvements in either school participation or learning outcomes. Six of these studies are briefly summarized below:

---

[10] This is mainly because the transfers are also provided to infra-marginal households who would have sent their children to school regardless of the CCT. On the other hand, note also that CCT's are broad social protection programs that aim to achieve several goals other than improving education and are perhaps by design not as cost-effective at improving education outcomes.

[11] These experimental findings are consistent with those from well-identified studies using difference-in-difference methods to study the impact of new school construction on enrollment (Duflo 2001) or on providing subsidized transport to improve access to school (Muralidharan and Prakash 2016).

[12] For instance, the Indian state of Rajasthan recently took a policy decision towards such a consolidation, though there has been no evaluation to date of its impacts.

Glewwe, Kremer, and Moulin (2009) conduct an RCT evaluating the provision of free textbooks to school children in Kenya and find that this had no impact on either student attendance or learning outcomes. Sabarwal, Evans, and Marshak (2014) experimentally evaluate a similar program in Sierra Leone and again find no impact on attendance or learning outcomes. Das et al. (2013) present results from an experimental evaluation of the provision of a block grant to schools in the Indian state of Andhra Pradesh (that was mainly used to buy books and materials) and find no impacts on test scores after two years of the program. Mbiti et al. (2016) conduct an experimental evaluation of a school grant program in Tanzania that was again mainly used for textbooks and learning materials and also find no impacts on test scores after two years of such grants. Borkum, He, and Linden (2013) experimentally evaluate a program in the Indian state of Karnataka that provided schools with libraries (consisting of providing a collection of books and a librarian rather than physical construction of a library – which makes the intervention closer to one that provided books rather than infrastructure) and find no impact of the program on learning outcomes. Finally, Pradhan et al. (2013) evaluate the impact of a community engagement program to improve schooling in Indonesia where they include a treatment arm that provided a block grant as a benchmark against which to study the other interventions, and find that the provision of the block grant had no impact on learning outcomes. While the reasons for the zero effect may vary across specific studies and contexts (see discussion in section 4.2), the breadth of the evidence above spanning Africa, South Asia, and Southeast Asia suggests quite strongly that simply providing more resources to schools is unlikely to improve learning outcomes.

The broad theme that that simply providing inputs to schools and students may have limited impacts on learning comes is corroborated by evidence on a different class of educational inputs – namely computers. A striking example of this is provided by Cristia et al. (2012) and Beuermann et al. (2015) who conduct a large experimental evaluation of the "One Laptop per Child" program in Peru, and find that even though the program led to a sharp increase in the fraction of children with access to a computer, it had no impact on learning outcomes. In contrast, interventions that effectively use technology to improve pedagogy typically have positive impacts on test scores (see section 3.3).

Another critical input into education is teachers and teacher salaries account for the majority of education spending in most countries. The key research questions of interest are studying the impact of teacher quantity and teacher quality on education outcomes. In practice, the number of teachers hired is determined by the average class-size norms that school systems try to achieve, and so the key research question of interest is the impact of class size (also referred to as pupil-teacher ratio) on learning outcomes.

The experimental evidence on class size in developing countries is limited because it is not easy to randomly assign civil-service teachers across schools. However, there is indirect experimental evidence on the impact of class size from multiple studies. The best evidence comes from the "Extra Teacher Project" analyzed in Duflo, Dupas and Kremer (2011, and 2015). The project conducted an RCT in Kenya that randomly assigned some schools an extra contract teacher. This extra teacher was assigned to grade 1 and used to reduce class size by roughly 50% (halving the pupil-teacher ratio from roughly 80:1 to 40:1). Further, half the students in first grade in treatment schools were randomly assigned to classes taught by current (civil-service) teachers, and the other half were assigned to classes taught by contract teachers. For the purpose of identifying the impact of the pupil-teacher ratio on student learning the classes taught by the current (civil service) teacher can be compared to those taught by same type of teacher in the control schools, which have much larger pupil-teacher ratios. Duflo, Dupas, Kremer (2015) report that although this reduction in class size led to higher test scores (about 0.09σ), this increase was not statistically significant.

Another piece of indirect evidence is provided by Banerjee et al. (2007) who conduct an experimental evaluation of a remedial education program where students with low test scores were "pulled out" of class for remedial instruction for a few hours each day. While the test scores of the students who received this remedial instruction went up significantly, the authors find no impact on test scores of students who remained in the original classes with smaller class sizes, suggesting that smaller class sizes may have limited impact on improving learning outcomes.

On teacher quality, there is limited experimental evidence on the impact of teacher training programs though panel-data based studies typically find no correlation between possession of a teacher training credential and measures of teacher value-added (Muralidharan 2013). The experimental study that most closely resembles a study of teacher training is Muralidharan and Sundararaman (2010) who study the impact of providing teachers with detailed written diagnostic feedback on the performance of their students on external assessments along with suggestions for more effective teaching, and find that there was no impact whatsoever of the program on learning outcomes (not only were the point estimates of impact zero, but the distributions of test scores in treatment and control groups were nearly identical). In interpreting these results, the authors present complementary evidence to suggest that the zero impact was not because the content of the diagnostic feedback was not useful, but rather that teachers in government-run public schools did not have the motivation or professional incentives to use this feedback and revise their teaching practices.[13]

[13] Specifically, they show that the extent to which teachers report that the feedback reports had useful content was not correlated with student test score gains in the schools that only received the feedback reports, but that it

Finally, teacher salaries are often considered an important source of teacher motivation and many global education advocates recommend raising teacher salaries to improve their motivation, and effort. Reflecting this thinking the Government of Indonesia passed an ambitious new teacher reform in 2005 that permanently doubled the base pay of teachers who passed a simple certification process. Using a large scale experiment that accelerated the doubling of pay for teachers in randomly selected schools, De Ree et al (2016) show that the program significantly improved teacher satisfaction with their income, reduced the incidence of teachers holding outside jobs, and reduced self-reported financial stress. However, despite these changes in measures of teacher well-being and satisfaction, there was no impact on either teacher effort or student learning.

In contrast with the mostly negative results summarized so far, one class of input-based interventions that have shown promising results are those that improve student health. The most striking example of this is the positive impact (both short and long-term) of school-based deworming treatments in settings with high rates of child worm infections. Miguel & Kremer (2004) provide experimental evidence on the impact of a school-based deworming program in Kenya and find that it led to a significant improvement in school attendance in treated schools. While they did not find impacts on test scores in the short run, a long-term follow-up study found significant long-term positive impacts from exposure to the program. Baird et al. (forthcoming) report results from the long-term follow up and find that ten years after the treatment, men who were eligible for the treatment as boys were more likely to be enrolled in primary school, worked 17% longer hours each week, and missed one fewer meal per week. Women who were eligible for the program as girls were more likely to have attended secondary school. The authors estimate that the annual internal rate of return of the deworming program was 32%, making it a highly cost effective program.

Overall, the experimental evidence on various categories of school inputs suggests that most of the "business as usual" patterns of expenditure seem to be rather ineffective at improving education outcomes. Why might this be the case? As I discuss further in sections 3.3 and 3.4, there is considerable evidence to suggest that providing traditional inputs to schools does not typically alleviate binding constraints (of pedagogy and governance) to improving learning outcomes in many low-income settings. Thus, better inputs *may* have a positive impact in settings where these other binding constraints have been alleviated, but appear to have more limited impact on learning outcomes in developing country settings where constraints in pedagogy and governance are more first order.

---

was significantly positively correlated with student test score gains in schools that received both feedback reports and performance-linked bonuses for teachers. Thus, when teachers were rewarded for improving student test scores it appears that they did make use of the diagnostic feedback reports, but not otherwise.

3.3 Pedagogy

While education inputs (infrastructure, teachers, and materials) typically account for the majority of education expenditure, a critical determinant of the extent to which these inputs translate into better learning outcomes is the pedagogy in the classroom. In other words, the "technology of instruction" is a key determinant of how inputs are translated into outcomes, and improvements in this "technology" may yield substantial improvements in outcomes. Researchers have only recently started to use experimental techniques for identifying the impacts of pedagogical innovations on learning outcomes, and there is a relative paucity of experimental evidence on the critical question of how best to structure classroom instruction and pedagogy to most effectively improve learning outcomes in developing countries. Nevertheless, progress has been made on better understanding these issues in the past decade, which in turn has provided important insights into the nature of the education production function, and the binding constraints that confound attempts to improve education outcomes in developing countries.

The most important insight on pedagogy that has been obtained in the past fifteen years of experimental research on education in developing countries is that education systems in developing countries are particularly poorly equipped to handle the pedagogical challenges posed by tens of millions of first-generation learners entering formal schooling. In particular, the curricula and teaching practices that may have been optimal at a time when education was more limited have not been adapted to deal with the challenge posed by increasing heterogeneity in classroom preparedness. Also, as Muralidharan and Zieleniak (2014) show, the variance in student learning levels within a cohort increases over time as they pass through the school system.

Thus, designing effective pedagogical strategies for handling variation in academic preparation across students in a classroom is a fundamental challenge for effective teaching and learning, and the experimental evidence over the past fifteen years has consistently found large positive effects of interventions that aim to address this challenge. We highlight three categories of interventions: supplementary teaching at the right level (TaRL), classroom tracking, and individually-customized computer aided learning.

The pedagogical intervention with the most consistent evidence of strong positive effects is that of supplemental education that ignores the text book and instead focuses on teaching children "at the right level". In practice, this means that children who are unable to read are taught the alphabet; those who can read alphabets are taught to put them together to read words; those who can read words are taught to read sentences and so on. These programs (many of them designed by *Pratham* – an education non-profit based in India) are typically delivered by young women who have completed secondary school or high-school who do not have any formal teacher training credentials and who are paid very modest stipends.

Several RCT's of *Pratham's* "Teaching at the Right Level" (TaRL) program have been carried out over the past decade across different locations and implementation models and the effects of these interventions have been positive in many settings ranging from urban locations in Western India (Banerjee et al. 2007) to rural North India (Banerjee et al. 2010, 2016; Duflo et al. 2015). Further evidence in favor of this approach is provided by Lakshminarayana et al. (2013), who study the impact of a program run by a different non-profit (the Naandi Foundation) that recruited community volunteers to provide remedial education to children in a randomly selected set of villages in the Indian state of Andhra Pradesh, and find that student test scores in program villages were 0.74σ higher than those in the comparison group after two years.

More recently, Duflo et al. (2015), and Banerjee et al. (2016) present results from multiple RCT's conducted across several Indian states (Bihar, Uttarakhand, Haryana, and Uttar Pradesh) in partnership with *Pratham* to evaluate the impact of different models of implementing the TaRL approach in public schools. Overall, they find support for the hypothesis that Pratham's instructional approach, which focuses on teaching children at a level that matches their level of learning, can significantly improve learning outcomes. However, they find considerable variation in the effectiveness of different implementation models. They find that implementing the pedagogy in summer camps that are held outside of normal school hours or in dedicated learning camps in the school (at which point children were grouped by their level of academic preparation as opposed to by grade level), was highly effective in raising test scores.[14] However, they found that it was more difficult to achieve large gains under implementation models that attempted to incorporate this pedagogy into the teaching by regular teachers during the school day.

For example, the first attempt to scale up the TaRL in public schools in Bihar and in Uttarakhand consisted of training regular teachers in the TaRL pedagogy and providing instructional materials to support implementation, but these programs did not lead to any improvement in learning outcomes. In Bihar, a variant of the program that also had volunteers providing supplemental instruction outside the school using the TaRL method improved language and math scores by 0.11σ and 0.13σ. However, in Uttarakhand, even providing an extra volunteer did not improve test scores because the volunteers were provided within schools and were used by teachers to implement regular as opposed to TaRL pedagogy. Recognizing this challenge, the design of the TaRL in program Haryana relied on schools

[14] Note however that the model with in-school learning camps was much more effective overall because student attendance at the summer camps was limited, whereas student attendance rates were much higher at the in-school learning camps (see the further discussion in the conclusion regarding scaling up successful interventions).

dedicating an hour a day for the TaRL pedagogy, where students in primary schools were reorganized on the basis of learning levels as opposed to the grade they were enrolled in. Duflo et al. (2015) evaluate this implementation model and find that it improved reading scores by 0.15σ. Finally, the most successful implementation model was that of "learning camps" conducted within government schools in Uttar Pradesh, where Pratham volunteers essentially took over the running of schools for 50 days (through 5 10-day camps including one in the summer) and achieved remarkably large gains of 0.7σ in both math and language.

The authors interpret their findings as suggesting that the remedial pedagogy was successful, but that it was difficult to get teachers to implement new curriculums during school hours, and that successfully scaling up remedial pedagogy within an existing schooling system can be challenging because teachers are focused on completing the syllabus prescribed in the textbook. Successful models required either dedicated reorganizing of school time committed to by the government (as in Haryana) or the taking over of school time by volunteers focused on implementing the TaRL model (as in Uttar Pradesh).

A second way of reducing the variance in student preparedness within a classroom is to "track" students into streams within a grade as a function of their preparedness. While proponents argue that tracking would benefit all students by allowing teachers to better tailor their level of instruction, opponents are usually concerned that students who are tracked to "lower" level classrooms may suffer further from negative peer effects and from stereotyping and loss of self-esteem, which may place them on a permanently lower trajectory of learning.

Duflo, Dupas, and Kremer (2011) provide important evidence on this question by experimentally evaluating a program in Kenya that tracked students in first grade into two classrooms based on their baseline test scores. They found that students in tracked schools had significantly higher test scores (0.18σ) than students in non-tracked schools, and that they continued to score 0.18σ higher even one year after the tracking program ended, suggesting longer-lasting impacts than those found in many other education interventions. They also found positive impacts for students at all quartiles of the initial test score distribution and could not reject that students who started out above the median score gained the same as those below the median; Additionally, lower-achieving students gained knowledge in basic skills, while higher-achieving students gained knowledge in more advanced skills, suggesting that teachers tailored their classes to the achievement level of their students. Finally, since students just below and just above the median baseline score were otherwise similar, but experienced a sharp change in the mean test score of their peers, the authors are able to use this regression discontinuity method to show that tracking did not cause adverse peer effects in this setting.

A third way of differentiating instruction for students at different levels of preparation is computer-adaptive learning. While there are several reasons to be optimistic about the *potential* for technology to improve learning outcomes (including the ability to overcome limitations in teacher knowledge, tailor instruction to students' needs, and provide real-time feedback to students), the evidence to date on the impact of interventions that simply provide computer hardware suggests zero to negative impacts on test scores (Barrera-Osorio and Linden 2009; Cristia et al. 2012; Beuermann et al. 2015; Malamud and Pop-Eleches 2011). On the other hand, interventions that focus on using technology for better pedagogy have typically found more positive results.

Banerjee et al. (2007) evaluate a computer-aided learning (CAL) program that consisted of math games that were designed to emphasize basic competencies in the official math curriculum in 2 cities in Western India, and find large gains in test scores at the end of one and two years of the program (0.35σ and 0.47σ respectively). Further, a series of experimental evaluations of computer-aided learning (CAL) in China have found modest positive impacts on learning (in the range of 0.12σ to 0.25σ). These studies include Lai et al. (2011) who study a CAL program in schools for migrant children in Beijing; Yang et al. (2013) who study a CAL program in 3 provinces in China (Shaanxi, Qinghai, and Beijini) in schools for socio-economically disadvantaged students. However, the CAL programs in the experiments in China used technology to reinforce grade-appropriate content and did not feature extensive individual customization, and the CAL program in India featured children sharing computer time, which may have limited the ability of the system to provide individual customization.

Recent evidence on the potentially dramatic benefits of individually-customized CAL programs is provided by Muralidharan, Singh, and Ganimian (2016) who experimentally evaluate a computer-aided learning program (called *Mindspark*) that was explicitly designed to customize pedagogy to the right level of students in grades 6 to 9 in New Delhi, India. The *Mindspark* program was developed iteratively over many years by a leading education diagnostic testing firm in India and featured a question bank of over 40,000 individual questions calibrated finely to students existing achievement levels using over 2 million observations of student item responses. Further, the *Mindspark* system analyzes patterns of student responses to the screening test and algorithmically identifies areas of conceptual misunderstanding. Thus, the computer-adaptive learning system was able to finely calibrate students' baseline competencies and tailor academic content to that level. Further, the system dynamically adjusts the content that students are working on at any given point of time based on the response to the previous question. This ability to customize instruction may be especially important in post-primary grades where the variation of student ability is likely to be even larger than in primary grades.

Using detailed question-level data, Muralidharan, Singh, and Ganimian (2016) report five results. First, students in this setting are several grade-levels behind their grade-appropriate standard, and this gap grows by grade. Second, in the control group, students in the lower half of the baseline test-score distribution in a grade show *no improvements* in test scores during the school year, suggesting that the level of the "business as usual" curriculum may be too high for students in the lower-end of the baseline test score distribution to learn anything effectively. Third, regularly attending the Mindspark program led to large increases in student test scores of 0.6σ and 0.4σ in math and Hindi (language) test scores respectively; these represent a three-fold increase in math and a three-and-a-half-fold increase in Hindi test score value-added relative to non-participants. Fourth, consistent with the promise of customized instruction, the treatment effects are equally large at all levels of absolute baseline scores. Fifth, because the "business as usual" rate of learning for academically weaker students is close to zero, their relative improvement from the technology-aided curriculum is much greater.

These results suggest that curricular mismatch may limit the extent to which time in secondary school translates into additional learning in developing countries (especially for students with low foundational skills), and that CAL programs designed to adjust to the level of the student can sharply improve learning outcomes for all students in such settings.

The main insight from the evidence summarized in this section is that a key binding constraint to converting inputs into learning outcomes in developing countries may be the fact that "business as usual" instructional practices that follow the textbook and curriculum are not at the correct level for the majority of students in developing country education systems. Attempts to address this problem that have been successful to date include supplemental remedial instruction, tracking, and customized computer-adaptive learning tools.

Thus, investing in designing effective pedagogy to handle the large variation in student preparation in developing country classrooms is likely to yield considerable positive returns in terms of improved education quality in these settings. Further, in addition to designing and evaluating technical solutions for better pedagogy, considerable additional efforts are required to embed these improvements in systems to improve pedagogy at scale. As the evidence in Banerjee et al (2016) shows, it is not easy to change default teaching practices to incorporate evidence on new and more effective pedagogy. These are all areas where much more experimentation and research is needed.

3.4 Governance

A fourth critical determinant of education outcomes in addition to household demand, the provision of inputs, and the details of classroom pedagogy is the overall governance of the education system. I use the term "governance" broadly to include goal-setting for the education system, motivating service providers for delivering on these goals and holding them accountable if they do not, and the quality of management of schools and school systems.

One striking indicator of weak governance in schools in developing countries is the high rate of teacher absence. Chaudhury et al. (2006) present results from a multi-country study where enumerators made unannounced visits to public schools to measure teacher attendance and activity, and report an average teacher absence rate of 19%, with teacher absence rates of 25% in India and 27% in Uganda. Muralidharan et al (2016) present more recent estimates from a nationally-representative panel survey that revisited the rural villages surveyed by Chaudhury et al. (2006), and find only a modest reduction in teacher absence rates, from 26.3% to 23.7%. They also calculate that the fiscal cost of teacher absence is $1.5 billion *each year*, highlighting the large costs of poor governance in education. Another example of weak governance is direct corruption in education spending where large amounts of funds meant for schools simply do not reach them. For instance, Reinnikka and Svensson (2004), show that 87% of central government funds allocated to a school capitation grant (for non-wage expenditures) in Uganda never reached the schools, and that the median school in their representative sample had not received *any* of the funds.

A considerable body of experimental evidence has been accumulated over the past decade on the impact of interventions to improve governance and management of education systems in developing countries. The broad summary of this literature is that while there is variation in the effectiveness of individual interventions, there is enough evidence to suggest that there are highly cost-effective interventions that can substantially improve governance, if there is political willingness to do so. In other words, the evidence suggests that *technical* solutions exist for improving governance, but that scaling these up may be difficult without the political desire to do so. I briefly summarize four categories of interventions to improve school governance: decentralizing more authority to communities for school management, modifying teacher contractual structure, performance-linked pay for teachers, and private management/vouchers.

The theory of change behind decentralizing school management is that providing more authority to communities to hold schools and teachers accountable for performance will improve teacher effort and student learning outcomes. However, the evidence from five different RCT's of such policy changes in different settings suggest that this approach may not be very effective. Banerjee et al. (2010) present experimental evidence on a program in rural North India that tried to empower communities to oversee their schools by making them more aware of their rights over schools and found no impact on learning outcomes. Pradhan et al (2014) find limited effects of enhancing community participation in school governance in Indonesia. Beasley and and Huillery (2014) experimentally evaluate a program providing grants to school committees to encourage parental participation in school management in Niger and find no impacts on test scores. Finally, Lassibille et al. (2010) and Glewwe and Maiga (2011) both present experimental evaluations of the AGEMAD program in Madagascar that aimed to strengthen school management at the district, sub-district, school and teacher levels, and find no impact on student test scores of these interventions.

A likely explanation for these results is that communities in practice do not have much authority over teachers, who are typically civil service employees and also typically much more educated (and hence powerful) than the residents of the communities that they serve (see Kingdon 2011 for a discussion of how the large 'social distance' between teachers and communities makes it difficult for the latter to hold teachers accountable). For instance, Pradhan et al. (2014) find that though having elections for school committees did not improve outcomes on its own, it did improve learning outcomes when implemented as part of a treatment that also provided formal linkages to the village council through joint planning meetings. The authors argue that this was because the village council was more powerful and that the linkage provided greater legitimacy to co-sponsored activities of the village council and the school committee. In contrast, enhanced community participation alone did not provide the school committee with enough power to impact learning.

Further evidence in support of the idea that a key constraint to the effectiveness of school management committees is their authority over schools is provided by Duflo, Dupas, and Kremer (2015). They find that training school management committees to evaluate the performance of contract teachers and to have inputs into the renewals of contract teacher contracts had a significant positive impact on the performance of the contract teachers and on student test scores. Thus, in cases where the committees had direct authority over the renewal of teacher employment contracts (and where the teachers belonged to the same community), they were able to make a difference. Thus, improving accountability of teachers is likely to matter, but not all policy attempts (like increasing community participation) may be effective at doing so.

A second way of improving teacher accountability and performance is to modify the structure of employment contracts so that the renewal of employment is contingent on measures of performance. Of course, most teachers in both developed and developing countries are employed in the public sector, and public sector employment contracts typically feature lifetime tenure after a very short probationary period (if any). However, in recent years many developing countries have started to employ new teachers on short-term renewable contracts. Contract teachers comprise a third of public-school teachers across twelve countries in Africa (Bourdon et al. 2010) and their share among all public-school teachers in India grew from 6 percent in 2003 to 30 percent in 2010 (Muralidharan et al. 2016).

There are two notable experimental studies on the impact of contract teachers. First, Duflo, Dupas, and Kremer (2015) conduct an RCT of the impact of providing randomly-selected schools in Kenya with an extra contract teacher, who was assigned to grade one and used to reduce class size in first grade by half (with students randomly assigned to either the section with a contract teacher or the one with a regular teacher). They find that class-size reductions with regular teachers did not have a significant impact on test scores, but that students with reduced class sizes who also had a contract teacher scored 0.29σ higher than those in control schools. They also find that holding class size constant, students taught by contract teachers scored significantly higher than those taught by civil-service teachers.

Second, Muralidharan and Sundararaman (2013) conduct an RCT of contract teachers of the impact of providing randomly-selected schools in the Indian state of Andhra Pradesh with an extra contract teacher, where the schools were free to assign the teacher as they wished (since the schools featured multi-grade teaching, the optimal use of the teacher would vary considerably across schools and it would be difficult to ensure fidelity to a within-school randomization design). They find that students in schools with the extra contract teacher had significantly higher test scores after two years of the program, and show using several non-experimental techniques that the contract teachers were at least as effective as regular teachers at improving test scores. Both studies find that contract teachers had significantly lower absence rates than regular teachers, and also find that the absence rates of regular teachers increased in schools with an extra contract teacher, suggesting that the estimated effects may be a lower bound on the true effect of an extra contract teacher.

Contract teachers tend to be different from civil service teachers in many ways (they are typically less educated, less likely to have formal teacher training credentials, more likely to be from the local community, and typically paid much lower salaries) and neither of the studies above (or any other study) can isolate the impact of just changing the contractual structure of employment holding all other factors constant. However, since most of the other differences would suggest that contract teachers may be less effective than regular teachers (such as being less educated, less trained, and paid much lower salaries), the evidence from the two experiments suggests that the renewable nature of the contracts may have contributed to the greater accountability of the contract teachers.

A third way of improving teacher effort and motivation that has been well studied is the idea of introducing performance-linked pay for teachers.[15] There are four noteworthy

---

[15] There are several reasons why default compensation systems for teachers have little or no link to performance. These include difficulties in measuring productivity of individual teachers, as well as concerns that linking pay to performance on measurable attributes of a job will lead to diversion of effort away from socially valuable tasks that may not be as well measured (Holmstrom and Milgrom 1991, Baker 1992). Nevertheless, the demonstrated low levels of teacher effort in developing countries (manifested by high rates of absence) have led both policy makers and researchers to consider the possibility that introducing performance-linked pay for teachers may improve outcomes.

experimental studies on this topic. First, Muralidharan and Sundararaman (2011) conduct an RCT of a program in the Indian state of Andhra Pradesh that paid teachers bonuses on the basis of the average improvement in test scores of their students. They find significant improvements in test scores of students in treated schools after two years (of 0.27σ and 0.17σ in math and language) and find no evidence of any negative effects. Students in treated schools did better on both "mechanical" and "conceptual" components of the test, where the former were designed to reflect questions in the text book (that could be "taught to") while the latter were designed to reflect deeper understanding of the materials. Students in treated schools also did better on science and social studies tests (for which there were no incentives) suggesting positive spillovers from improvements in math and science.

Second, Muralidharan (2012) presents evidence from a 5-year long follow-up of the original experiment (where a randomly selected subset of the original schools saw the continuation of the performance-pay program for five years) and reports strong positive effects for the cohort whose teachers were eligible for performance pay for five years (their test scores were 0.54σ and 0.35σ higher in math and language). The study also finds that though group and individual teacher incentives appeared equally effective in the short-run, the individual incentives significantly outperformed the group incentives at the end of five years.

Third, Glewwe, Ilias, and Kremer (2010) conducted an experimental evaluation of a teacher incentive program in Kenya that provided school-level group incentives using prizes for high-achieving schools, and find that students in treatment schools did score better on high-stakes tests but not on low-stakes tests, and also that these gains dissipated after the incentive program was over. They interpret their results as suggesting that teacher incentives may not be effective as a strategy for promoting long-term learning. Nevertheless, there are two important caveats. The first is that we now know that all interventions appear to have significantly high rates of test-score decay (see Andrabi, Das, Khwaja, and Zajonc 2011) and that there may be important long-term gains in human capital even when test score gains decay (Chetty et al. 2011). Second, the group-nature of the incentive program (across 12 teachers) may have induced free riding and weakened the incentives faced by individual teachers (as seen in Muralidharan 2012).

Fourth, Duflo, Hanna, and Ryan (2012) present evidence from a program in the Indian state of Rajasthan, that paid teachers based on the number of days they attend work as opposed to a flat salary and find that the program led to a halving of teacher absence rates (from 42% to 21%) and significant increases in student test scores (by 0.17σ). The intervention combined both better monitoring (teacher attendance was verified with photos with time-date stamps) and better incentives (since were paid based on days of attendance). Using a structural model identified by non-linear sections of the pay-off schedule, the authors show that the improvement in attendance is mainly attributable to the improvements in incentives as opposed to just increased monitoring by itself.

All the programs studied above featured relatively small bonuses that averaged less than 10% of monthly pay. The large positive effects from even modest amounts of pay linked to performance are particularly striking when compared with the finding of *zero* impact on student learning from an unconditional *doubling* of teacher pay in Indonesia. Taken together, these results suggest that even modest changes to compensation structure to reward teachers on the basis of objective measures of performance (such as attendance or increases in student test scores) can generate substantial improvements in learning outcomes at a fraction of the cost of a 'business as usual' expansion in education spending. However, not all performance pay programs are likely to be effective, so it is quite important to design the bonus formulae well and to make sure that these designs reflect insights from economic theory (see the discussion in section 5.1).

The final class of governance interventions that has attracted considerable policy and research attention is the idea of combining public funding for education with competition across public and private producers of education, through voucher-based models where parents get to choose schools (public, private, or non-profit) and the government directly reimburses the school. The promise of such voucher and choice based reforms is that private management may be more effective than traditional public school management and that giving parents more choice across schools (as opposed to limiting them to publicly provided schooling options) would increase competition and accountability across all schools.

School choice is a controversial subject, and experiments are particularly important in testing the relative effectiveness of public and private schools, because cross-sectional differences in test scores are highly likely to be confounded by omitted variables. There are two sets of well-identified studies of school voucher programs in developing countries that defrayed the cost of attending private schools. Angrist et al. (2002) and Angrist et al. (2006) study the short and medium term effects the PACES program in Colombia that provided vouchers (allocated by lottery) to students to attend private schools, and find that voucher winners scored significantly better both three and seven years after receiving the voucher. However, the PACES program also allowed vouchers to be topped up by parents (to attend a better school than they could have afforded without a voucher), and required students to maintain minimum academic standards to continue receiving the voucher. Thus while the results point to the effectiveness of the PACES program, the estimates reflect a combination of private school productivity, additional education spending, and student incentives.

Muralidharan and Sundararaman (2015) present experimental evidence on the impact of a school-choice program in the Indian state of Andhra Pradesh that featured a unique two-stage randomization of the offer of a voucher (across villages as well as students). The design created a set of control *villages* that allows the authors to experimentally evaluate both the individual impacts of school choice (using the student-level lottery) as well as its aggregate effects including the spillovers on non-applicants and students who start out in private schools (using the village-level lottery). At the end of two and four years of the school choice program, they find no difference between the test scores of lottery winners and losers on the two main subjects of Telugu (the native language of Andhra Pradesh) and math, suggesting that the large cross-sectional test-score differences in these subjects across public and private schools (of 0.65 standard deviations) mostly reflect omitted variables.

However, they find that private schools spend significantly less instructional time on Telugu (40% less) and math (32% less) than public schools, and instead spend more time on English, and science and social studies. They also taught a third language, Hindi, which was not taught in public primary schools. When they conduct tests in these additional subjects after four years of the voucher program they find small positive effects of winning the voucher on English (0.12 standard deviations; p-value = 0.098), and science and social studies (0.08 standard deviations; p-value = 0.16), and large, positive effects on Hindi (0.55 standard deviations; p-value < 0.001). Further, the annual cost per student in the public-school system is over three times the mean cost per student in the private schools in the sample.

Thus, on the one hand, private schools were clearly *more productive* than public schools (they achieved similar results on the main subjects at much lower cost, and produced gains on an additional subject that was not taught in the public schools), but they were also *not more effective* at improving learning outcomes on the core subjects. The results suggest that private management may have the potential to deliver better learning outcomes at comparable costs, but there is no evidence yet that this is the case. Thus, a key open question for future research is to study the relative effectiveness of private and public management holding the spending per student constant.[16]

3.5 Summary of Evidence

As mentioned earlier, the review above does not aim to be a comprehensive review of all field experiments in education in developing countries (see Glewwe and Muralidharan 2016 for a recent summary). Rather, it aims to synthesize the consistent patterns in the evidence and highlight the most important general insights obtained from field experiments in education in developing countries in the past decade.

[16] The recently announced initiative by the Government of Liberia to launch the "Partnership Schools for Liberia" initiative provides a promising opportunity to answer this question.

The key messages from this evidence is that "business as usual" expansion of spending on school inputs (which is where the majority of education spending is allocated) may have only modest impacts on improving education outcomes. The main reason for this appears to be that the binding constraints to better performance of developing country education systems appear to be not inputs but rather: (a) outdated pedagogy that focuses on completing textbooks without accounting for the fact that millions of new first-generation learners may be considerably behind the levels assumed by the textbook; a problem that gets worse in older grades, and (b) weak governance with poor accountability for teachers and other front-line service providers. Thus, these appear to be the most important areas to focus attention on, in addition to designing effective demand-side interventions.

Nevertheless, as the discussion of evidence shows, not all interventions to improve demand, pedagogy, or governance are equally effective, which underscores the need for ongoing high-quality evaluations of these initiatives. The next sections aim to provide guidelines for young researchers on how to effective design and implement such evaluations in education in developing countries.

4 Limitations of Field Experiments and Strategies for Mitigating them

In this section, I provide a discussion of the important limitations of field experiments. Many of these limitations apply to almost *all empirical research that tries to identify causal relationships*, and should not be seen as weaknesses of experimental methods in particular. But it is important to be clear about what problems experiments do and do not solve, and doing so can improve the quality of policy-relevant inference made from individual studies, and may also help guide future research in ways that mitigate these challenges. The goal of discussing these limitations here is to (a) provide the necessary nuance and caveats in interpreting the results discussed in section 3, and (b) to motivate the discussion in Section 5 where I describe ways of addressing these limitations through better design and data collection.

4.1 Production Function versus Policy Parameters

The discussion in section 2.3 highlighted the value of experimentally varying $X_{i,t}$ in estimating the causal impact of $X_{i,t}$ on education outcomes. Note however, that even random assignment of $X_{i,t}$ may not yield the production function parameter β outlined in Eq. (1). This is because the production function parameter β is a partial derivative ($\partial T_{i,t}/\partial X_{i,t}$) holding *other inputs constant*. In practice, other inputs at the school or household level may endogenously respond to exogenous changes in $X_{i,t}$, and the estimated parameter should therefore be more accurately interpreted as a policy parameter, which is a total derivative ($dT_{i,t}/dX_{i,t}$) that accounts for re-optimization by agents in response to an exogenous change in $X_{i,t}$.

The extent to which an experimental estimate reflects re-optimization will depend critically on the duration of the study. A clear illustration is provided by Das et al. (2013), who study a randomly-assigned school grant program in India over a two-year period and find significant positive effects on test scores at the end of the first year, but find no effect in the second year even though the grant was provided again in the second year, and was spent on very similar items in both years (books, school supplies, and classroom learning materials). They show that the most likely explanation for this result is that household spending on books and school supplies did not change across treatment and schools in the first year (when the school grant was unanticipated), but that households in treatment schools sharply cut back their spending on these categories in the second year (when the school grant was anticipated and could be accounted for in household decision making), and that this reduction offset around 80% of the per-student value of the grant.

The authors therefore argue that the "first year" effect of the program is more likely to represent the "production function" effect of providing the school grant (since other factors did not have time to adjust), whereas the "second year" effect is closer to the "policy parameter" (which reflects household re-optimization). The example highlights the value of measuring as many intermediate inputs as possible to have a better idea about the mechanisms of program impact. However, in practice, it will be difficult to measure *all* possible intermediate inputs, and the extent to which they may have changed in response to the exogenously-varied treatment. Thus, it is perhaps most accurate to interpret the "causal estimate" of β from experimental studies as the "policy effect" of $X_{i,t}$ at the point when the outcomes are measured.

Note that this limitation is also present for non-experimental methods, and is therefore not a criticism of experiments per se. But it an important limitation to highlight because experimental estimates are often implicitly interpreted as representing production function parameters based on Eq. (1). This may well be true over short time periods where other agents may not have re-optimized behavior, but it is (a) difficult to confirm that this is true on every dimension of potential behavior modification, and (b) much less likely to be true over longer horizons.[17] One advantage of well-identified evaluations using large administrative data sets (based on regression discontinuity designs for example) is that it may be possible to observe the policy effects at longer time horizons at much lower marginal cost than in experimental studies (since the cost of conducting follow up surveys on experimental samples can be quite large, and the challenge of differential attrition grows over time). A good example of this is provided by Bharadwaj et al (2013) who can measure the impact of early childhood

[17] While the discussion may suggest that experimental estimates may be lower bounds of production-function parameters and upper bounds of policy parameters, this need not be true if the unmeasured inputs are complements to the experimental intervention as opposed to substitutes (as was the case in Das et al 2013).

interventions several years later using administrative data in both Chile and Norway. Longer-term follow ups of experimental interventions are relatively rare, but should be a higher priority for funders and researchers.

This discussion also unifies the approaches outlined in sections 2.3 and 2.4. One advantage of using the approach in Attanasio (2015) as outlined in section 2.4 of evaluating experiments through the lens of a constrained household choice problem is that re-optimizing behavior is directly built into the problem framework as opposed to being an afterthought. It therefore forces the researcher to be disciplined about how the intervention affects either the constraints (production function, resources), beliefs/attention (as with knowledge interventions), or household preferences (as with interventions that may affect intra-household bargaining) and to interpret the effects through this unified lens.

4.2 Interpreting zero effects

A second challenge in conducting inference from experimental studies is that of interpreting zero effects. In theory, this should simply mean that the estimate of β in Eq. 1 is zero and that the marginal impact of increasing $X_{i,t}$ is insignificantly different from zero. In practice, however it is important to distinguish between five different interpretations of a zero result. These include (a) poor implementation of the intervention, including corruption or administrative failures, (b) substitution away of other inputs by agents (including governments, schools, teachers, and households) in response to the treatment, (c) positive effects on some sub-populations but not on others, leading to an average effect that is not significantly different from zero, (d) absence of complementary inputs/reforms that may be needed for the intervention to be effective, and (e) a true zero effect for all students. Note that reasons (c), (d), and (e) are consistent with the interpretation that the marginal impact of increasing $X_{i,t}$ on outcomes is zero in a production function sense, but reasons (a) and (b) are not. Further, the distinction between (c), (d), and (e) also matters for policy because the policy implication of (e) would be to not prioritizing increasing $X_{i,t}$, whereas that of (c) would be to provide it to the sub-group where it was effective, and that of (d) would be to increase $X_{i,t}$ as long as the complementary input is also increased.

These possibilities are illustrated across four different randomized evaluations of the impact of providing books and materials to students. Each of the four studies find zero average impacts of providing books and materials, but point to different possible reasons for the zero effects. Sabarwal et al (2014) find no impact on test scores from the provision of textbooks to schools in Sierra Leone and attribute this to the fact that schools actually stored the textbooks instead of distributing them to students (which is a form of poor implementation). Das et al (2013) described above also find no net impact on test scores from the provision of a school grant (that was mostly spent on books and materials) in India, but attribute it to households offsetting the intervention by reducing their own spending on these inputs. Glewwe et al (2009) also find no impact on test scores from providing textbooks to students in Kenya. But they find positive impacts on students with the highest baseline test scores and suggest that their results are consistent with the fact that the majority of children could not read the English language text books to begin with, and thus could not benefit from the textbooks (whereas those who could read *did* benefit).

Finally, Mbiti et al. (2016) also find no impact on test scores from the provision of a large capitation grant to schools in Tanzania (the largest item that the grant was spent on was textbooks). However, their study was explicitly designed to test for complementarities with teacher effort (which was boosted by a separate intervention that paid teachers bonuses based on student performance) using a cross-cutting design with a sample size large enough to test for complementarities, and they find that the interaction effect of the school grant and teacher performance pay was significantly positive. In other words, the school grant on its own had no impact, but had a significant impact when provided in conjunction with a teacher performance pay intervention. Thus, it is likely that the performance pay treatment contributed to teachers making more effective use of the additional materials, but it is also true that having the materials allowed teachers to significantly improve student outcomes relative to teachers who only increased effort due to performance-linked pay.

The larger point here is that each of these experiments with zero results are useful results in and of themselves, and yield an important policy conclusion that the marginal impact of providing books and learning materials to students may be very low on their own. On the other hand, the fact that four papers with the same result point to four different reasons for this non-impact suggest that a "black box" experiment on its own may yield limited insights into the nature of the education production function and the true binding constraints to learning.

More generally, it is really important and useful to document *why* interventions that funders and implementers spend so much time and money on have no impact (if they do not). Papers finding zero impact are very important, but will typically contribute more to learning if accompanied by careful analysis of intermediate variables to better understand and describe which parts of the posited causal chain of impact worked and which ones broke down. Thus, it is good practice for researchers to think ex ante about how they would interpret a zero effect, and to collect as much data as possible on implementation quality as well as intermediate inputs and processes to enable better interpretation of finding no effects of a program.

4.3 External Validity

Perhaps the most widely discussed limitation of experiments is that of the external validity of their results beyond the specific setting where they are carried out (Cartwright 2007; Deaton

2010). The formal way of thinking about this problem is to recognize that though the random assignment ensures that un-observables are distributed identically across treatment and control groups and that the treatment is not correlated with these un-observables, the estimated program effects are for not for the treatment *alone*, but rather for the treatment *interacted* with the unobservable characteristics in the study sample. If these unobservable characteristics vary between the study sample and the universe to which we seek to extrapolate the findings to, then the estimated treatment effects may not be valid because the interactions may change.

There are several variants of this concern that are worth spelling out distinctly, because the strategies for mitigating them are different. There are at least four limitations to generalizing experimental results even in the *same context* where the experiment was carried out. I discuss these first before discussing external validity *across* contexts.

4.3.1 External Validity in the Same Context: Representativeness of Study Samples

First, there is a concern of external validity even in the context of the evaluation because most experiments are carried out within a universe of schools that agree to participate in the experiment. If these schools are different from those who do not agree to participate (perhaps their leadership is more open to trying out new ideas), then the results might have limited external validity (Heckman and Smith 1995).[18] Most experimental studies in education to do not pay enough attention to this issue, and it is not that difficult to do so (see section 5.2.1).

4.3.2 External Validity in the Same Context: Implementer Heterogeneity

Second, a further concern with external validity even in the same context comes from the fact that many RCT's in education evaluate interventions implemented by committed implementation partners (often highly motivated NGO's). Thus experimental estimates of programs implemented by NGO's may not translate if the same program is implemented by the government (as shown in Bold et al. 2013). The differences in the results they report between government and NGO implementation of a contract teacher program largely reflects the fact that the program itself was very poorly implemented by the government. So, it does not negate the results found under NGO-implementation, but it does highlight that programs are not just an "intervention", but rather an intervention *and* an implementation protocol. This is not a problem per se, but suggests that evaluations of NGO-led implementations should be

[18] A variant of this concern is seen in the US charter-school literature where well-identified estimates are only available on the causal impact of over-subscribed charter schools (which are likely to be the higher quality ones) as opposed to the universe of charter schools, which is the policy parameter of interest (unless the over-subscribed schools are able to expand without compromising quality and the schools that are not over-subscribed shut down).

seen as efficacy trials and not effectiveness trials.[19] It also suggests that when successful NGO-implemented interventions are being scaled up, there may be a strong case for conducting further RCT's at larger units of implementation and when implemented by the entity that will eventually scale up the intervention (typically a government).

4.3.3 External Validity in the Same Context: Varying intervention details

Third, even if a government wishes to use experimental results in a given context to guide policy in the same context, it is extremely unlikely that the policy chosen will be exactly the same as the one evaluated. For instance, even if a CCT is found to have a positive impact, the value of the CCT may be changed later for political or budgetary reasons. Similarly, even if a teacher performance pay program or a student incentive program is found to be effective, a policy maker would care about the elasticity of the outcome of interest to the magnitude of the incentives in order to better calibrate the value of the incentives. These questions are harder to answer within the context of an experiment because it is often politically and administratively difficult to vary the magnitudes of such incentives within an actual program. Further, experimentally estimating elasticities often requires sacrificing power or a larger sample. While it may be possible to do this, a more promising approach may be to combine experimental methods with structural modeling to allow more credible out of sample predictions than either of the two approaches could on their own.

Good examples of this are Todd and Wolpin (2006), and Attanasio, Meghir, and Santiago (2012) who combine structural models of school participation with observed impacts of the PROGRESA CCT program to enable predictions of program impact under alternative values of the cash transfer. Another good example is Duflo, Hanna, and Ryan (2012) who use the non-linearities in the compensation schedule of an experimentally evaluated teacher incentive program to identify parameters in a dynamic model of teacher labor supply and use the model to estimate cost-minimizing compensation policies to achieve a desired level of teacher attendance. However, these additions of structural models to enable out of sample predictions have mostly been done ex post and were not designed ex ante into the study, which may have limited the extent to which the experiment could be used to identify parameters in the structural model of interest. Future experimental research in education is likely to have greater

[19] These terms are standard in the medical literature and refer to the difference between impacts under high-quality implementation that is closely monitored (efficacy trial), and impacts under typical implementation that allows for typical patient behavior including non-compliance with dosage frequency and complementary instructions (effectiveness trial).

impact if the ex-ante design of the study includes careful thinking about the model that the experiment can be used to identify, and plans its data collection strategy accordingly.[20]

In practice, the skill set required to run high-quality field experiments is considerably different from that required to specify and estimate structural models. Thus, there are likely to be considerable benefits to forming teams of researchers with complementary skills across designing and running field experiments and structural modeling from the outset to conceptualize experiments from the dual perspectives outlined in sections 2.3 and 2.4 and to design data collection strategies accordingly.

4.3.4 External Validity in the Same Context: Political Economy

A fourth and final concern regarding external validity in the same context is that experiments cannot typically capture the general equilibrium effects (both political and economic) that may accompany attempts to scale up successful smaller scale experiments. In the words of Acemoglu (2010), “Political economy refers to the fact that the feasible set of interventions is often determined by political factors, and large counterfactuals will induce political responses from various actors and interest groups. General equilibrium and political economy considerations are important because partial equilibrium estimates that ignore responses from both sources will not give the appropriate answer to counterfactual exercises”.

A good example of this is the case contract teachers, where existing experimental and non-experimental evidence suggest that locally-hired contract teachers who are less educated, less trained, and paid much lower salaries than civil-service teachers are at least as effective (if not more) at improving learning outcomes in rural primary schools in both Kenya and India (Duflo et al. 2014; Bold et al. 2013; Muralidharan and Sundararaman 2013). Thus, expanding the use of contract teachers on the current margin would appear to be a very promising and cost-effective policy for improving education outcomes in developing countries. Nevertheless, scaling up contract teacher programs has been difficult politically because forward looking officials are aware that hiring a large number of contract teachers will lead to them getting unionized and creating political pressure to get “regularized” as civil-service teachers, which is very difficult for politicians to ignore.

---

[20] For instance, household surveys that accompany school-level treatments are often collected in repeated cross-section samples. This makes sense if the goal is to characterize the *average* difference in household inputs across treatment and control groups, because cross-sectional heterogeneity is typically much greater than over time heterogeneity within the same household, and a repeated cross-section that covers more households will yield more precise estimates of these averages. However, a structural modeling exercise that aims to understand household decision making in response to a school-level treatment will typically benefit from repeated surveys of a smaller sample of households to better model the dynamics of household choice in response to realizations of information from their child’s school over time.

This is a problem that is difficult to solve empirically with an experiment, but the discussion above highlights the importance of treating positive results from an experimental evaluation of an intervention as just one of many inputs into the policy-making process. Finding positive technical results from an RCT of an intervention can be a good starting point for considering the administrative and political challenges of scaling up and designing implementation protocols that take these into account, but it would be naïve to recommend "scale ups" based on RCT evidence alone. It is perhaps not a coincidence that the leading example of policy scale up based on RCT evidence is de-worming, which is administratively easy and politically costless. On the other hand, other interventions with robust evidence of positive effects (like the use of contract teachers) have been much more difficult to scale up. In such cases, the experimental evidence is best treated as the starting point for a more informed policy conversation.[21]

4.3.5 External Validity Concerns across contexts

Obtaining external validity across contexts is even more challenging, given that the unobserved covariates (that would interact with the treatment of interest to produce the average treatment effect) are likely to be different across contexts. This problem is well known among academic researchers, but is often under-stated in "systemic reviews" that compare interventions and estimates across contexts (see Pritchett and Sandefur 2013 for a discussion). There is no good experimental solution to this problem beyond conducting more studies and gathering more evidence by replicating evaluations of similar (if not 'identical') interventions in many settings. The problem of external validity from well-identified individual studies is now receiving more formal attention (see Dehejia, Pop-Eleches, and Samii 2015 for an approach that derives an external validity function based on matching on observed covariates), and is likely to an area of fruitful research because the analytical standards that have been external validity in the past decade have been much weaker than those applied to internal validity.

Despite the many successful field experiments in education in developing countries in the past decade, the overall experimental research agenda in education in developing countries is still at an early stage. Some pieces of evidence seem quite robust across several contexts (such as the lack of impact of providing books and materials to students), and others have been replicated in multiple sites in the same country (such as "teaching at the right level" across states in India), but most other interventions do not have enough replications across contexts to enable confident claims of their impacts across contexts. A further problem is that there is considerable variation in the details of interventions implemented and evaluated across

[21] See Muralidharan (2013) and Muralidharan (2016) for an example of a policy proposal that takes the results from evaluations of contract teacher programs seriously, and accounts for administrative and political economy considerations in recommending feasible policy approaches that are consistent with the evidence.

settings, which makes generalization even more challenging (see the discussion in Glewwe and Muralidharan 2016 on this).

Thus, to the extent that donors and development agencies seek summaries of evidence (as seen by the eight summaries written in the last 2 years), attempts to calculate comparative cost effectiveness of interventions conducted across several contexts should be interpreted cautiously (Dhaliwal et al. 2013). In trying to learn across contexts from the experimental literature to date, it may be more appropriate to focus on "principles" that have been validated in multiple settings rather than the "point estimates" of specific studies. The summary of the evidence presented in section 3 reflects this approach.

5. Conducting Field Experiments in Education in Developing Countries

There are several excellent resources for researchers and practitioners wanting to design, implement, and analyze field experiments including Duflo, Kremer, and Glennerster (2007), Gerber and Green (2012) and Glennerster and Takavarasha (2013). Nevertheless, there is a considerable amount of tacit knowledge that is accumulated by researchers through practice in a particular area (such as education) that is often not available easily, and the goal of this section is to synthesize some of this knowledge and offer it as a resource for researchers and practitioners wanting to design, conduct, analyze, and interpret field experiments in education in developing countries.[22]

5.1. Design

Getting the design right is the core of a good experiment, and requires considerable thought and attention up front. The importance of investing in this upfront thinking is perhaps best illustrated with an anecdote about doing economics research in the 1960's that I heard from a very senior researcher. He mentioned how a graduate student doing empirical work would often just be able to run *one* regression for an entire dissertation since computer time would have to be booked many months in advance and punch cards carefully prepared in advance of being able to run the one regression. The implication was that all the thinking had to be done in advance, and had to checked and re-checked multiple times because you only got to run the regression once and would lose months of time if things did not work out.

Conducting field experiments is similar in many ways in that you typically only get to run the experiment once, and so it is essential to get all the thinking done upfront. This may seem obvious, but the best experimental papers have a deceptive "easiness" about them that

[22] There is nevertheless a high degree of "learning by doing" in conducting field experiments. It is therefore common practice (and highly advisable) for young researchers to initially work on teams led by senior researchers (including as research assistants) to obtain the tacit knowledge that is best obtained by practice.

typically hides the large amount of advance thinking that goes into a well-done experiment. Experiments are unforgiving of mistakes in that it is typically not possible to change the design of treatments or the overall experiment once it is underway. Thus, it is essential to obtain multiple sources of feedback before commencing a field experiment, and the best experimental designs often get iterated several times at the design stage. The discussion below covers some of the main considerations in designing a good field experiment in education and is organized around two main topics – intervention design and experiment design.

5.1.1. Intervention Design: What to look for?

While experimental methods can help with credible estimation of the causal impact of interventions, it is important to ensure that adequate thought is given to determining whether the intervention being studied is worth evaluating and the extent to which the findings of a study generate more generalizable knowledge (especially given the non-trivial time and effort costs of setting up an RCT). In particular, it is not uncommon to see competently implemented and analyzed experiments in education, where the underlying intervention is rather ad hoc and not adequately theorized, which limits what we learn from the evaluation. In this section, I offer some (personal) guidelines for informing the decision on whether an intervention is worth evaluating experimentally.

Three useful questions to ask before deciding if it makes sense to conduct an experimental evaluation of an education intervention are: First, is there genuine uncertainty about the impact of the intervention. Second, is the intervention addressing a well-understood supply or demand side deficiency and *designed well enough* to address this deficiency. Third, are governments spending large amounts of money doing things whose effectiveness we do not understand well enough (e.g. infrastructure, class size reductions, teacher training, teacher salary increases, school feeding programs, school grants)?

A good rule of thumb for identifying whether an intervention is worth studying experimentally is not just to want to test "if a program works" but to test ideas where there is genuine uncertainty and controversy regarding their impact. Experimental evaluations of education interventions are typically more influential when there are compelling theoretical arguments both in favor of and against an intervention, and where the answer is essentially an empirical question. Examples include the impacts of student tracking (Duflo, Dupas, and Kremer 2011), the impacts of linking teacher pay to measures of effort (Duflo, Hanna, and Ryan 2012) or gains in student test scores (Muralidharan and Sundararaman 2011), and the impacts of school choice and private schools (Muralidharan and Sundararaman 2015). These studies are characterized not just by studying the impact of the intervention on education outcomes, but by paying serious attention to the hypothesized negative impacts and taking care to measure these potential negative impacts to reach a more nuanced and complete understanding of the impact of the intervention.

In the case of tracking, the concern of opponents of tracking regarding negative peer effects on students assigned to the “lower performing” track was tested by Duflo, Dupas, and Kremer (2011) by combining an experimental evaluation of tracking with a regression-discontinuity based evaluation of whether there were negative peer effects. In the case of the teacher performance-pay program evaluated by Muralidharan and Sundararaman (2011), the authors designed the program to mitigate against some of the known concerns of teacher performance-pay programs and tested for others (such as teaching to the test and potential negative effects on non-incentive subjects). In the case of school choice, a key concern has been the possibility of negative spillovers on students who did not apply for the voucher or who were in private schools to begin with and were exposed to lower-achieving peers who transferred in to these private schools with vouchers. The study addresses these issues with a two-stage randomization design that allowed the researchers to quantify these spillovers.

A second class of interventions worth studying are those where there is a well-hypothesized theory of change from an intervention to outcomes and the program is not necessarily controversial, but where the program is not being implemented (typically due to the lack of a champion within the government or budgetary constraints). In these cases, a high-quality evaluation accompanied by cost-effectiveness calculations can be a very useful contribution to research and policy by helping to make the case for allocating public funds for expanding the program (if effective) or for not doing so (if found ineffective).

A good example is the provision of de-worming tablets to school children, which was found to be a much more cost-effective way of increasing school attendance than other spending on other student inputs (Miguel and Kremer 2004, Dhaliwal et al. 2013), and has since resulted in scaled up deworming programs in many parts of the world. A second example is the growing body of evidence on the effectiveness of “Teaching at the Right Level” (see Banerjee et al. 2015), which is increasingly leading to governments being interested in scaling up the core idea.

Both these ideas seem obvious ex post, but did not receive much policy maker attention before the research results. In the case of de-worming, the cost of the intervention was trivial, but school health programs (under which such an initiative would typically have to be implemented) would often suffer from coordination failures across health and education ministries, without clear ownership. In the case of “Teaching at the Right Level”, the idea that children are not learning because they are far behind where the default level of classroom instruction simply does not seem to cross the mind of many policy-making elites because the situation is outside most of their own experiences.[23] Thus, high-quality evaluations of innovative interventions can be a very useful catalyst in making non-obvious constraints to education quality salient to policy makers, and obtaining policy consideration for cost-effective solutions to alleviate these constraints.

A third category of evaluations worth conducting are those where the policy/program being evaluated is something that governments spend a lot of money on. These evaluations can be very useful even if the researcher has ex ante reasons to believe that the intervention is not well designed or may not be effective because the money will be spent on the program anyway, and it is very useful for policy to understand whether the program was effective. Further, to the extent that the program (as designed) reflected conventional wisdom that it would have a positive impact, the evaluation could shed light not just on the "program" as implemented but also on the hypotheses underlying the design of the program.

A good example of such an evaluation is de Ree et al. (2016) who study the impact of an unconditional doubling of teacher pay in Indonesia. While we did have a prior expectation that this doubling may not have much impact on student learning (or at least that the same money could have been much better spent), the evaluation was still worth conducting because (a) the policy was very expensive, and (b) many education advocates believed that increasing teacher pay would improve teacher motivation, effort, and student learning.

5.1.2. Intervention Design: What to avoid?

One common mistake is to rush into an RCT before the intervention being studied has been adequately piloted, codified, and stabilized. If the intervention is being modified during the study, it is difficult to interpret the findings. Thus, it is advisable for programs and interventions to be "standardized" and "easily replicable" before embarking on an RCT. A related challenge occurs with RCT's of "composite" interventions that include components that are not easy to codify, which makes it difficult to interpret the results of an RCT. Note that a "composite" intervention per se is not a problem since it is often possible that there are complementarities across components of the package and the "package" may be the intervention that we need to

[23] Indeed, one plausible reason for why well-intentioned education interventions may have limited impact on learning outcomes is that policies and programs are typically designed by elites whose experiential understanding of education may not correspond to that of the representative student. For instance, on a field visit to schools in Tanzania, I saw that several children would typically share one textbook. It is not difficult to imagine how this would be a highly salient fact for a visiting senior education policy official or foreign aid official, and result in a well-intentioned program to provide free textbooks that would take considerable financial and administrative resources to deliver. However, as the results in Glewwe et al. (2009) suggest, the provision of free textbooks may not alleviate the binding constraint for the average student in this setting – which is that they cannot read. Similarly, media discussions of education in countries such as India focus disproportionately on the issues relevant to the high end of the achievement distribution (who comprise the readership of newspapers) as opposed to issues relevant to the representative student.

evaluate. Rather, it is the inability to codify and replicate the "package" that limits the learning from an evaluation of composite interventions.

Another pitfall to be aware of is that of studying "gold-plated" interventions that have high unit costs. The risk in such a setting is that a high-quality (but high-cost) intervention gets evaluated and is found to have a substantial positive impact, but is difficult to scale up because of a lack of financial resources to sustain the program.[24] A further risk is that a diluted version of the high-cost prototype is scaled up (on the basis of the evaluation), but that this version may not have any impact. One option for imposing discipline in this regard is to not only have a "pure" control group (that does not get any additional intervention), but to have other comparison groups that are provided an equivalent amount of resources, which enables a direct cost-effectiveness comparison against reasonable policy alternatives. An example of this is provided by the Andhra Pradesh Randomized Evaluation Studies (APRESt) where the impact of teacher performance pay was evaluated not just against a pure control group, but against comparison groups that received an equivalent valued school grant or extra contract teacher.[25]

A third category of evaluations to be careful of are cases where the details of the intervention have not been well designed, in which case even an experimental evaluation may not contribute much to learning. This point is best illustrated by the example of teacher performance pay. A common way for such policies to come into being is that a policy-maker may decide to implement some form of teacher performance pay, and administrators then design a particular formula for paying teachers performance-based bonuses. However, the optimal design of a performance pay system is a non-trivial problem and in many cases the formulae designed by administrators are likely to have important design flaws that are likely to limit the effectiveness of performance pay. For instance, many simple formulae reward teachers based on the number of students who pass a performance threshold (like passing a test), which does not provide incentives for teachers to continue improving student learning above the threshold, or far below the threshold (see Neal and Schanzenbach 2009). In such a case, an evaluation that does not a find a significant impact of performance pay is not very useful beyond being able to say that the specific program as implemented was not effective.

The design of teacher and student incentives is a good example of a case where economists can add value not just in terms of conducting a “well identified evaluation”, but in using

[24] One manifestation of this is the phenomenon of donor-financed pilot projects being abandoned once the donor financing is over. Of course, these programs typically do not have credible impact evaluations to inform the decision on whether developing country governments should continue to spend on them out of their own budgets. But scaling up of high unit-cost interventions would be difficult even with positive evidence of impact.

[25] Blattman and Niehaus (2014) make a similar point with regard to evaluations of anti-poverty programs in general, proposing that the benchmark should not just be a pure control group, but rather an unconditional cash transfer that is equal in value to the full cost of the program being evaluated.

theoretical first principles to design better variants of the intervention. There is an extensive theoretical literature on incentive design that covers topics ranging from piece rates versus tournaments (Lazear and Rosen 1981; Green and Stokey 1983); linear versus non-linear incentives (Holmstrom and Milgrom 1987); group versus individual incentives (Itoh 1991; Kandel and Lazear 1992; Kandori 1992) and a corresponding empirical literature with examples from outside education. There is also a theoretical literature on optimal design of teacher performance pay systems (see Barlevy and Neal 2012) that also has implications for the optimal design of student incentive programs. There are several papers on teacher/student incentives that do not pay attention to this theoretical literature and are weaker as a result.

So the main advice from this discussion is to make sure to pay at least as much attention to the *design of the intervention* as to the design of the evaluation. Further, when researchers have influence over the design of the intervention, the design will almost always benefit from understanding the relevant theoretical literature to inform the decisions and trade-offs that will have to be made in finalizing the intervention. More generally, any field experiment in education will benefit from researchers starting the study by clarifying the research question, and asking themselves what they would learn about the world and about the intervention from different potential values of the treatment effect.

5.1.3. Experiment Design: Unit of Randomization

As discussed earlier, the "modular" nature of many education interventions makes it feasible to randomize them at relatively small unit levels, including at the student, classroom, and school levels. The main consideration in determining the unit of randomization is the trade-off between statistical power[26] for a given budget (which is higher for lower units of randomization) and the possibility of spillovers, which may bias the experiment and negate the power gains from randomizing at a lower level.

A striking example of this trade-off is seen in the difference between impact evaluations of deworming that randomized the treatment at the student level and those that randomized the treatment at the school level, such as Miguel and Kremer (2004). While the former typically found very limited impacts of deworming on education outcomes, Miguel and Kremer (2004) found large positive impacts on treated students, as well as spillovers from treated to non-treated students. The spillovers would under-state the impact of deworming in studies with student-level randomization in two ways. First, it would under-estimate the impact on the treated students (because the non-treated students who served as the control group also benefited from the intervention). Second, it would not count the impact on non-treated

[26] Note that this chapter does not spend too much time on generic issues of experimental design such as power calculations, for which there are many other references. Rather the focus is on design choices that are especially relevant to research in education.

students who also benefited from the intervention. Of course, randomizing at the school level significantly increases the sample size required for adequate power, and correspondingly increase the cost of the study. On the other hand, it is not clear that the gain in power from student-level randomization is worth it if it biases the treatment effect itself.

Note however, that the case of de-worming may be an outlier in terms of the extent of spillovers across students. Several studies within the "Rural Education Action Program" (REAP) have utilized student-level randomization within schools in China to study the impacts of interventions ranging from peer-tutoring to student incentives. Berry (2013) uses student-level randomization to study the impact of different combinations of student and parent incentives.

Nevertheless, given that students interact with each other every day in the classroom, it is difficult to credibly claim that there would not be any spillovers (especially for treatments implemented in schools as opposed to households), which may limit the extent to which we can learn from such experiments. One important exception is cases where the treatment is at the school level and control students are not in treated schools, as happens in cases where students receive vouchers (or charter-school admission) by lottery and do not interact with control students during the course of the school day. But overall, student-level randomization designs should be used with caution and limited to situations where the focus of the intervention is outside the school setting (such as households or after-school programs).

A less extreme set of concerns applies to designs that randomize at the grade or classroom level as opposed to the school level. Again, the main reason for doing so is power and cost. Several prominent studies have used classroom-level randomization designs to study the impact of remedial instruction (Banerjee et al. 2007), tracking of students according to initial ability (Duflo et al. 2011), and comparing contract and regular teachers (Duflo et al. 2015). The spillover concerns are less severe in this case because most instructional activity as well as peer interactions between students happens at the classroom level and not across classrooms (which is where the spillovers would have to happen).

However, there is still a concern that interventions that provide a significant increase in resources to some classrooms, may lead a head-teacher to offset some of the impact of the treatment by reallocating some other resources to control classrooms. This could happen either due to norms of fairness within the school or because an optimizing head-teacher would reallocate resources to equalize their marginal product across classrooms.[27] Such behavioral responses could contaminate the experiment and the inference.

This is why my personal preference has been to randomize at the school-level to the extent possible for studying interventions ranging from teacher performance pay, across the board teacher salary increases, provision of school grants, provision of diagnostic feedback to schools, and also the provision of an extra contract teacher. The overall logic of this approach is that a

[27] For instance, the model in Lazear (2006) predicts that more disruptive students will optimally be assigned to smaller classrooms.

policy maker can typically target resources at the school level, but cannot easily control how those resources are allocated within the school by an optimizing head-teacher. Thus, the policy-relevant parameter of interest in these cases is the impact of a school-level provision of an intervention. The limitation of this approach of course is that the samples need to be much larger and the studies cost more.

One rule of thumb for choosing between school and grade/classroom level randomization may be to consider the size of the school. In smaller schools (as is the case in rural India where most primary schools have less than 100 students across 5 grades and the modal school has only 2 teachers teaching in multi-grade classroom settings), it is difficult to convincingly argue that a within-school randomization protocol that assigns a program to just some grades will be adhered to without re-adjustment. On the other hand, when schools are much larger and have several hundred students and dedicated teachers in each grade (as is the case in many African settings), the fidelity of within-school experimental protocols may be more reasonable to assume because teachers spend all their time with one grade as opposed to teaching multiple grades at the same time. Further, the costs of school-level randomization may be prohibitive in settings with such large schools.

Overall, I believe that it is less of a problem for an intervention to be targeted at specific grades within a school (as opposed to the entire school) as long as the control group is the same grade in a *different* school as opposed to other students within the same school. In such a setting one may still worry about spillovers attenuating the treatment (if resources are diverted to other non-treated grades in treatment schools), but at least the control group will not be contaminated. Studies that use within-school controls should have the burden of proof for demonstrating that the controls were not contaminated by spillovers by the time that the endline measurements are conducted.

5.1.4 Experiment Design: Power and Baselines

Beyond standard discussions of power calculations and sample size calculations, the following three strategies can increase power in education interventions. First, given the autocorrelation in student test scores over time, it usually makes sense to conduct student-level baseline tests and include these test scores as controls in estimates of treatment effects. Second, additional gains in precision (and power) can be obtained by also controlling for school-grade-subject level baseline means in the treatment effect regressions, in addition to the student's own lagged test scores (Altonji and Mansfield 2014). Finally, stratifying the randomization within geographic/administrative regions can further increase power because the stratum fixed effects will absorb unobserved spatial variation and reduce the sum of squared residuals, which in turn will increase power and precision (Muralidharan and Sundararaman 2011; de Ree et al. 2016).

Nevertheless, it can also make sense to not do a baseline and it is possible to conduct high-quality experimental studies without a baseline round of testing. Reasons to not do a baseline include: risk management, time, and budget. In settings where there is a reasonable risk that the intervention may not be implemented, it may be prudent to not do a baseline, but to conduct a randomization for the intervention roll-out, and only expend time and effort on an evaluation if the intervention was actually implemented, and the randomization was successfully adhered to. Such an approach may also be needed in cases where there isn't adequate time to put in place field teams and raise research funding for the baseline. In such cases, it may be adequate to use administrative data for conducting the randomization (even if the administrative data is only at the school level and not the student level).

Such an approach may be especially useful when researchers are working with governments, where the risk of non-implementation of the intervention and non-compliance with a randomization protocol are higher. In such cases, it may make sense for researchers (especially risk-averse junior researchers with tighter resource constraints) to initially focus efforts on ensuring randomization and implementation of the intervention (and to push for a larger scale of implementation if possible), and to compensate for the lack of a baseline by increasing the sample size of the endline. Note that under government implementation, the sample size (and power) is usually not constrained by the intervention budget but by the evaluation budget and sample.

Finally, another reason for young researchers to consider this approach is that funders are much more likely to view your proposal favorably if you can demonstrate that the implementing agency has demonstrably adhered to the randomization and experimental roll out protocol. So an optimal approach may be to apply for smaller amounts of pilot funding to travel to the location of the study, interact with implementation partners, influence the intervention design (to the extent possible), collect administrative data, conduct the randomization, and ensure that the randomization protocols are followed during implementation. This can then be followed up by applying for funding for the endline, at which stage the implementation risks of the project would have been lowered considerably.

This is also a good place to discuss the issue of cost effectiveness of study designs. While many of the points made in this section may seem like they can only be addressed by senior researchers with larger budgets, the discussion above highlights that it is often possible to considerably reduce the cost of an RCT. In particular, studies that use administrative data on student outcomes can be especially cost effective to conduct, and should be a high priority for both researchers and funders going forward.

5.1.5 Experiment Design: Cross-cutting designs and interactions

Many experimental studies on education in developing countries aim to compare the relative effectiveness of different interventions on education outcomes in the same setting.

One commonly used approach to reducing the cost of conducting multiple studies in a given setting is to employ "cross-cutting" or "factorial" designs. In its simplest form, this usually involves a 2 by 2 matrix of 4 treatment cells with one group receiving neither treatment (the control group), a second group receiving just the first treatment (T1), a third receiving just the second treatment (T2), and a fourth receiving both (T1 and T2). The logic of such cross-cutting designs is best expressed by Kremer (2003) who notes that: "Conducting a series of evaluations in the same area allows considerable cost savings. Since data collection is the most costly element of these evaluations, cross-cutting the sample reduces costs dramatically. *This tactic can be problematic, however, if there are significant interactions across programs*."

Several highly influential papers have taken this approach to randomization and a particularly good example is Duflo, Dupas, and Kremer (2011) that was conducted as part of an "Extra Teacher Project" that aimed to study several important questions simultaneously in the same setting including (a) the impact of class size reduction, (b) the impact of contract versus regular teachers, (c) the impact of tracking students by initial test scores, and (d) the impact of community monitoring of schools. The project managed to study all 4 of these questions with a sample of just 210 schools by using a cross-cutting design with 70 control schools and 140 schools assigned to a combination of the treatments. The key to this approach generating consistent estimates is to pre-commit to a view that interactions among treatments are not important. This allows the researchers to treat the treatment effects as additive and makes it possible to double-count the schools with multiple treatments under each treatment (to increase power within a given measurement budget).

But in practice, interactions *are* likely to be important, and there is risk of overstating the effects of individual treatments if there are complementarities across treatments. Mbiti et al. (2016) show this in the context of an experimental evaluation of school grants and teacher incentives in Tanzania, that was explicitly powered to test for complementarities (and committed to this in a pre-analysis plan). They show that (a) each treatment was insignificant on its own when interactions are accounted for, (b) that the interactions between treatments were positive and significant, and that (c) ignoring the interactions would lead to an over-statement of the individual treatment effects.

Thus, the concern, about cross-cutting designs expressed in Kremer (2003) are likely to be salient in practice. The quote from Kremer (2003) highlights that this was and is a well-known issue. But it made sense for early experiments to use cross-cutting designs and ignore interactions at a time when evaluation budgets were much tighter, because this allowed adequate power for the first order questions.

However, as we generate evidence that interactions do matter (and are detectable with designs that treat them as equally important as the main effect and allocate adequate sample size to detect them), the assumptions underlying cross-cutting designs that ignore interactions appear less tenable. Since evaluation budgets are growing (with programs increasingly being expected to set aside funds for evaluation), it may make sense to prefer cleaner designs with single treatments and direct comparisons of treatments without being confounded by interactions. It may still make sense to ignore interactions in some settings, but doing so should be justified and documented in a pre-analysis plan. Note that this does not imply that every component of treatments should be broken down and tested separately. Many treatments of interest are "composite" by design and should be evaluated that way. But such an approach does not assume that the interactions are zero, which is what cross-cutting designs do (see Mbiti et al. 2016 for a more extended discussion on the implications for experiment design and hypothesis testing).

5.2 Sampling, Randomization, and Implementation

5.2.1 Sampling and Representativeness

As mentioned in section 4.3.1, experimental papers in education often do not pay enough attention to the representativeness of the universe of schools, which can limit the external validity of the studies even in the context in which they are conducted. The ideal experimental protocol to address this problem is to try as hard as possible to first draw a representative sample of schools/students from the universe that the study is trying to extrapolate to, and then randomly assign these schools into treatment and control groups. Such a protocol provides much more external validity than studies carried out in a "convenience sample" of schools and allows policy makers to be more confident that the experimental estimates apply to the relevant universe of interest. Examples of such an approach include Muralidharan and Sundararaman (2011, 2013) in the Indian state of Andhra Pradesh, de Ree et al. (2016) in Indonesia, and Mbiti et al. (2016) in Tanzania. Each of these studies featured random assignment in near-representative samples that allowed the results to be credibly extrapolated to populations of ~80M (Andhra Pradesh), ~200M (Indonesia), and ~45M (Tanzania).

While this may not always be possible for reasons of cost and logistics, experimental studies should at least discuss their sampling procedure in more detail (which is often not done) and show tables comparing the study sample and the universe of interest on key observable characteristics (similar to tables showing balance on observable characteristics across treatment and control units). Examples of such analysis are provided in Muralidharan, Niehaus, and Sukhtankar (2016) and Muralidharan, Singh, and Ganimian (2016).

5.2.2 Randomization

Readers are referred to papers dedicated to the subject of randomization procedures and trade-offs associated with them (Bruhn and MacKenzie 2009) and to the more formal treatment in the companion chapter by Athey and Imbens in this volume (Athey and Imbens 2016). The main practical point I want to make is that there is a strong case for stratification of randomization (especially by geographic units) to the extent possible. There are several advantages of doing so.

First, it almost always increases power. There is usually considerable spatial heterogeneity in unobservable characteristics (especially when carrying out studies in representative samples as recommended in the previous section). Thus, stratifying randomization at a low-level geographic unit that also corresponds to a unit of government administration (such as a district or even a sub-district) and analyzing the experimental results with stratum fixed effects will soak up a considerable amount of unexplained variation in the outcome variable, and reduce the sum of squared residuals in the regression estimating treatment effects – thereby increasing power (see Muralidharan and Sundararaman 2011, 2013, 2015; de Ree et al. 2016; Mbiti et al. 2016 for examples). Second, it insures against the risk of other programs being (randomly) implemented in some areas and not others. While this does not increase bias ex ante (since these other programs could be implemented anywhere), it reduces the risk of ex post contamination of treatment effects. Since other programs typically get implemented at district or sub-district levels (especially those by other non-governmental actors) stratifying and including geographic fixed effects allows the researcher to "net out" these effects. Third, it provides insurance against compliance or data collection problems in a small sub-set of schools. For instance, it is not uncommon for survey completion rates to be lower in more remote areas, or for treatments to not get implemented in some areas. Instead of having to assume (and justify) that these cases of non-compliance are random, it is often cleaner to just drop the stratum from the analysis (see de Ree et al. 2016 for an illustration).

5.2.3 Implementation and Follow Up

It is not uncommon for implementation partners to make changes to the program design or implementation protocols, without realizing that these may compromise the research design (or the interpretation of experimental findings). Hence, constant contact between researchers and implementation partners is essential to make sure that implementation is on track as intended (ideally). If changes are unavoidable then such regular contact can help ensure that (a) changes in implementation protocol do not compromise the evaluation design, and (b) changes in intervention design are clearly documented to enable an accurate description of the program as implemented, which in turn will allow for better interpretation of the findings. Thus, it is essential to budget plan for regular monitoring of implementation quality.

5.3. Data Collection

5.3.1 Outcomes

The main outcomes of education interventions are student participation (enrollment and attendance) and learning outcomes (typically measured by test scores). Attendance is best measured using unannounced visits to schools during the course of the study. If this is not feasible, say for budgetary reasons, another option is to collect attendance data from school records during the time of end-line testing (though these are less reliable). A final option is to use the attendance rate during the end-line tests.

Test scores are the most commonly used outcome measure for RCT's in education, but the standards for psychometric practice in the RCT literature are quite mixed.[28] The default method of measuring treatment effects in education is to express these effects in terms of standard deviations of normalized test scores (normalized relative to the control group). However, the tests used in many education studies are often not designed systematically, and details of test construction are often not reported (even in Appendices). This is problematic because measured treatment effects can be quite sensitive to the sampling of questions from the universe of feasible test questions. For instance, if the test is conducted at a level that is too difficult for most students (floor effects), then measured treatment effects will be zero even if there was a meaningful impact on learning at a level that was below the level at which the test was given (see Singh 2015 for a discussion of the point, and Muralidharan, Singh, and Ganimian 2016 for a demonstration of the importance of this issue in practice). I highlight some important principles of test design that education RCT's should aim to follow:

First, have tests with a wide distribution of test questions by difficulty. This is often not the case with tests that are designed for rapid assessment of basic learning like the ASER or Uwezo tests (which can have considerable ceiling effects) or grade-appropriate tests (which will typically have large floor effects in developing countries). So it is important that tests be piloted and researchers should ensure that raw test scores are well-distributed. Preferably, researchers should select items using Item Response Theory (IRT) and design the most effective tests for the expected ability levels of students. It is not very difficult to do this: testing in 3-4 schools provides an adequate sample size and involves only a few days of fieldwork. Existing routines in most statistical packages make it easy to generate "Item Characteristic Curves" (ICC's), and scaled test scores.

A further issue in test construction is to make them comparable across studies. Where possible, items should have a subset which allows them to be linked (through IRT) on a common global distribution of student learning levels. This will not only enable better comparison across studies (which is essential for cost-effectiveness calculations) but will also allow researchers to place the study sample in the context of the wider distribution of ability in the population. Since most RCTs are not carried out in representative samples, having such information in an Appendix to a paper will be a good practice. Having such linked tests is makes it easier to (a) characterize the "business as usual" evolution of learning levels in the control group, and (b) express treatment effects in terms of fractions/multiples of the learning in the

[28] In developing countries, the best work on measurement is probably that from LEAPS studies in Pakistan (Andrabi et al. 2011) but these are an exception.

control group over the same period.[29] Such data also makes it possible to characterize heterogeneity in treatment effects in a much richer way than is normally done in education RCT's (see Muralidharan, Singh, and Ganimian 2016 for an illustration).

Finally, the main threat to the validity of experimental estimates of the impact of education interventions is the possibility of differential attrition of students between treatment and control groups. High levels of differential attrition can severely compromise the validity of experimental estimates and it is therefore essential to take efforts to minimize the risk of attrition (especially differential attrition across treatment and control groups) during the data collection process (see Glennerster and Takavarasha 2013 for a detailed discussion).

5.3.2 Intermediate inputs, processes, and mechanisms

The minimum measurement needed for conducting an RCT in education is a set of outcome measures (typically test scores) collected at the end of the study period (assuming that the intervention was successfully randomized). However, the interpretation of experimental results is considerably enriched when accompanied by high-quality data on intermediate inputs, processes, and mechanisms. Opening up the "black box" of treatment effects with such data usually yields much more insight than simply reporting treatment effects (Kling, Ludwig, Congdon, and Mullainathan 2016).

Key intermediate variables to collect data on include school and household expenditure and time use. The importance of the first is illustrated by Das et al. (2013), which was discussed earlier. The importance of the second is illustrated by Muralidharan and Sundararaman (2015) who find in their study of school vouchers that there was no impact on math and native language test scores of winning a voucher and attending a private school. However, they also find that private schools spend much less instructional time on math and native language and use the time saved to teach other subjects (where they strongly outperform the public schools as may be expected). Thus, the inference on the relative productivity of public and private schools would have been incorrect if the differential patterns of time use had not been accounted for.

More generally, in addition to financial costs, it also important to consider all opportunity costs of an intervention (which is often not done). Consider the case of modifying curriculum by teaching new content. It is crucial to also specify *what is being replaced* in the existing curriculum to make way for the new additions and to test if there are negative impacts on learning of subjects that may have had their instructional time reduced to make way for the

[29] Note that there is one technical challenge in doing such a comparison. The estimated multiple will vary as a function of the test score persistence over time (see Andrabi et al. 2013, and Muralidharan 2012 for a discussion). One solution to this problem is to present ranges of treatment effects as a function of the persistence parameter (see de Ree et al. 2016 for an application of this approach).

new content. On the other hand, if the new materials are being taught over and above the existing content, it is important to price the opportunity cost of student and teacher time. This may be low or high, but needs to be accounted for.

A good example of the importance of accounting for time use in schools is provided by Linden (2008) who finds that a computer-enabled instruction program had positive effects on student test scores when offered as an after school supplementary program, but had negative impacts when it was used to substitute existing teaching activity. Thus, an important lesson for evaluations of education interventions is to account for all costs of the intervention - including *time* and financial costs, and being clear about whether the impact on test scores is coming from additional time on task (either home or school work) or from using existing time more efficiently (either by providing inputs that improve the marginal productivity of time in school, by organizing pedagogy more efficiently, or by reducing slack during the school day). Measuring intermediate inputs to the extent possible is key to enabling such analysis.

Other intermediate variables that can shed light on mechanisms include data on teacher attendance and teaching activity. However, these are difficult to measure precisely within typical research budgets because the former requires multiple unannounced visits for precision, and the latter require extended classroom observation time to meaningfully capture variation in teaching practices and detect changes in teaching practice. However, rich insights into classroom processes can be obtained when such measurement is done well and future research would do well to consider cost-effective tools for measuring classroom practice (Bruns et al. 2011; Araujo et al. forthcoming).

5.3.3 Long-term follow-up

An important limitation of experiments is that their prospective nature makes it difficult to obtain long-term outcomes without waiting for a long period of time. Thus, the majority of education experiments report outcomes within a few years of the program. Nevertheless, it is extremely important for both research and policy to be able to understand the long-term impact of programs, and some of the most influential studies in education and human development have been those that have tracked long-term outcomes including the Perry Pre-School study (Heckman et al. 2009) and the Jamaica home visitation study (Gertler et al. 2014). A good example of a more recent experimental intervention with longer-term follow up is provided by Baird et al (2016) who study the 10-year impacts of the de-worming program in Kenya studied by Miguel and Kremer (2004).[30]

[30] It is worth noting that from a policy perspective, important complementary evidence to Miguel and Kremer (2004) was provided by Bleakley (2007) who presented historical evidence on the long-term positive impacts of a mass deworming program in the US. The combination of short-term experimental evidence and long-term

In addition to long-term follow ups of short-duration interventions, a related issue is that of estimating the impacts of treatments that are continued for a long period of time. This is especially relevant for estimating the impacts of changing school-level policies because these changes will affect students for many years (potentially for as many years as they are in school). However, few studies manage to do this for a combination of budgetary and practical reasons. Some exceptions are Muralidharan and Sundararaman (2015) who study the impact of a school choice program after four years of exposure to treatment, and Muralidharan (2012) who studies the impact of teacher performance pay over five years of exposure to treatment.

While it is not easy for studies with more limited budgets to plan for either long-term follow up or long-term experimental exposure to a treatment, it can be extremely valuable to do so. In particular, funders and researchers should try to support long-term follow ups in cases where the short term effects are highly encouraging.

5.4. Analysis

There are several high-quality existing resources on analysis of experimental data (including Gerber and Green 2013, and Athey and Imbens 2016), and the reader is referred to these for a more formal treatment of the topic. This section will briefly highlight issues that are salient for the analysis of experiments in education and outlines a recommended set of analysis for papers in this area to follow, which is typically organized around main treatment effects, heterogeneity, mechanisms of impact, and cost effectiveness.

5.4.1 Main Treatment Effects

A typical estimating equation for studying the impact of receiving an education intervention takes the form:

$$T_{isjk}(Y_n) = \beta_0 + \beta_1 \cdot T_{isjk}(Y_0) + \beta_2 \cdot Treatment_i + \beta_{Z_i} \cdot Z_i + \beta_{X_i} \cdot X_i + \varepsilon_{isjk} \qquad (5.4.1)$$

where $T_{isjk}(Y_n)$ represents normalized test scores for student *i* in subject *s* in grade *j* and school *k*, at the end of *n* years of the experiment. Since test scores are highly correlated over time, it is standard to control for baseline test scores to increase the precision of estimates.[31] Including stratum (often geographic) fixed effects (Z_i) helps to absorb geographic variation and increase efficiency, and is needed to account for stratification of the randomization. The main

evidence using historical data (albeit less well-identified variation) provided greater confidence in the policy value of launching mass deworming programs.

[31] The default baseline score that is controlled for is the score on the same subject and student, but in cases where no baseline test was conducted in the same subject, it is still useful to control for the mean normalized test score across all subjects for which a baseline test was available for the same cohort, or if the cohort did not have a baseline, then the corresponding school-level mean for older cohorts can be included to increase precision (see de Ree et al. (2016) for an illustration).

estimate of interest is β_2, which provides an unbiased estimate of the impact of receiving the treatment, and it is standard to estimate β_2 both with and without controlling for household socioeconomic characteristics (X_i).

Since the treatment effect β_2 above is calculated relative to the control distribution (which is a standard normal), β_0 will typically be zero (or the omitted fixed effect) and has no cardinal meaning. It is standard to report β_2 separately for each subject tested, and to also report the mean treatment effect averaged across all subjects tested to present a summary statistic of impact. Such a summary statistic is especially useful in interpreting studies with positive effects on some subjects and not on others. Since such variation could simply reflect sampling variation (see discussion in the next section on heterogeneity), a summary statistic across subjects is useful to report.[32]

Eq. 5.4.1 represents the standard value-added model (VAM) that is the workhorse of the education literature. Note that this VAM does *not* use the intuitive "difference in difference" approach (where the dependent variable would be the difference between current and lagged test scores). This is because there is very strong evidence from several settings that test scores are not fully persistent over time. In other words, there is considerable decay in test scores over time and β_1 in Eq. 5.4.1 is typically estimated to be in the range of 0.3 – 0.7. The standard "difference in difference" specification imposes a restriction that β_1 in Eq. 5.4.1 equals 1, which is typically rejected in the data. Thus, imposing this restriction would lead to misspecification of the estimating equation and potentially biased estimates of β_2, which is why the default specification in this literature takes the form in Eq. 5.4.1. See the excellent discussion of the relevant issues in Andrabi et al. (2011).

Test score decay (or incomplete persistence) over time is typically not a problem for estimating short-term treatment effects in education experiments, because randomization ensures that mean baseline test scores are comparable across treatment and control schools. However, decay presents challenges when evaluating longer-term treatment effects. Specifically, the challenge is that the specification in Eq. 5.4.1 can be used to consistently estimate the *n*-year effect of the programs (with $T_{isjk}(Y_0)$ on the right-hand side), but not the '*n'th*' year effect (with $T_{isjk}(Y_{n-1})$ on the right-hand side) because $T_{isjk}(Y_{n-1})$ is a post-treatment outcome that will be correlated with the treatment indicator. The literature estimating experimental treatment effects in education therefore typically estimates only the *n*-year effect. However, over time, the "loss" of test scores due to decay will typically be higher in treatment schools since they start each year with higher test scores. See Muralidharan

[32] At the same time, it is also not clear that all subjects should be weighted equally, which is why it is good practice to report results both by subject and averaged across subjects (see Muralidharan and Sundararaman 2015 for a discussion).

(2012) for a discussion of the implications of the distinction between "gross" and "net" treatment effects for the evaluation of education interventions over a longer-period of time.[33]

Since the main threat to the validity of an experiment is attrition, the analysis of treatment effects should typically be preceded by a clear description of attrition across treatment and control groups, and a test of equality of attrition levels across treatment and control groups. It is also good practice to present the student and school-level correlates of attrition and to test whether the same model using observables can predict attrition in both the treatment and control groups (see Muralidharan and Sundararaman 2011 for an illustration). In cases, where some differential attrition is unavoidable, it is standard to include two kinds of robustness checks. The first is inverse-probability reweighting of observations to recover the distribution of students in the baseline (this is typically only valid if the observable correlates of attrition are similar across treatment and control groups). The second is to use bounding techniques (see Muralidharan and Sundararaman 2015 for an illustration).[34]

5.4.2 Heterogeneity

Heterogeneous treatment effects are typically estimated with linear interactions across household, school, and teacher covariates, and it is standard to report whether there are differential treatment effects across any of these covariates. Nevertheless, it is important to be cautious in interpreting the results of such analysis for at least two reasons. First, interacting a randomly assigned treatment with a non-randomly assigned covariate does not provide exogenous variation in the latter.[35] Second, in the absence of a pre-analysis plan with well-theorized reasons for heterogeneity along specific dimensions estimated heterogeneous effects could simply reflect sampling variation and inference should be corrected for multiple comparisons. One way to make such analysis credible would be to pre-specify such heterogeneity in advance (see Olken 2015 for a discussion).

[33] This discussion is mainly relevant for studies that use test scores as the main dependent variable, and is less relevant for longer-term studies that track employment and earnings outcomes.

[34] Of course, the best situation is one that has limited attrition to begin with. This should be a high priority for data collection efforts (see Glennerster and Takavarashan 2013 for further discussion on how to do so in practice).

[35] A good example is provided in Table 6B of Muralidharan and Sundararaman (2011). They find that teachers with lower base pay responded more to the teacher performance pay program that they study. These results may suggest that the magnitude of the performance pay mattered because the potential bonus (which was the same for all teachers) from a given level of improvement in student performance would have yielded a larger bonus (as a fraction of base pay) for teachers with lower base pay. However, teacher base pay is also strongly correlated with years of experience, and they find that teachers with fewer years of experience also respond better to the program. This is consistent with the possibility that younger teachers may respond better to any treatment since it may be easier for them to change their behavior, which is a completely different interpretation of the reason for heterogeneous treatment effects.

At the same time, it is also possible that some kinds of treatment heterogeneity that are not anticipated or pre-specified in advance but discovered ex post, may help to make sense of the overall experimental results. Good examples include Glewwe et al (2009) on the impact of providing free text books in Kenya, and Muralidharan and Sundararaman (2015) on the impact of school choice in India. In the first study, the authors did not specify that they expected to stronger effects on students at the top of baseline test score distribution. Nevertheless, finding this result ex post made sense because many of the students with lower baseline test scores were not able to read, which made it unlikely that they would benefit from the provision of a free text book. Similarly, the school choice experiment studied in Muralidharan and Sundararaman (2015) was not designed to test for heterogeneity by the medium of instruction of the schools that voucher-winning students chose to attend. But the finding (using instrumental variable techniques) that the impact of switching medium of instruction from Telugu (the native language) to English was negative for content subjects was consistent with other research and provided important nuance to understanding the overall experimental results. Thus, it is important to both report such results and to suitably caveat their interpretation.[36]

A particularly useful parameter along which to test for heterogeneity is the baseline student test score, which can be treated as a summary statistic of educational inputs that students had received up to the point when they enter the study. Educational interventions are also well suited to non-parametric analysis of heterogeneity, and doing so as a function of endline and baseline test score distributions can both be useful ways of characterizing the heterogeneity of program impacts. I describe each approach below.

The first is to estimate quantile treatment effects (defined for each quantile τ as: $\delta(\tau) = G_n^{-1}(\tau) - F_m^{-1}(\tau)$ where G_n and F_m represent the empirical distributions of the treatment and control distributions with *n* and *m* observations respectively), with bootstrapped confidence intervals for inference. Note that this does *not* plot the treatment effect at different quantiles (since student rank order is not preserved between the baseline and end line tests even within the same treatment group). It simply plots the gap at each percentile of the treatment and control distributions after the program and compares test scores across treatment and control groups at every percentile of the endline distribution. Such a plot is especially useful as a visual test of first-order stochastic dominance between treatment and control groups.

[36] As a template for such writing, see Muralidharan and Sundararaman (2015), who start their discussion of heterogeneity of impact by medium of instruction by noting that: "Our experiment was *not* designed to identify heterogeneous effects by school characteristics, but we report some suggestive results that are likely to be important for future research designed explicitly to study such heterogeneity."

A second way to show heterogeneity is to plot non-parametric treatment effects by percentile of *baseline* score, where the treatment and control endline distributions are plotted separately with the x-axis being the percentile of baseline score. This plot allows researchers to characterize the treatment effect as a function of where students were in the initial test score distribution (see Muralidharan and Sundararaman 2011 for a detailed illustration of these two types of analysis). However, this can only be done for cohorts for which baseline data exist.

5.4.3 Mechanisms

Mechanisms of impact are typically shown by comparing data on school and household inputs such as spending, and time use across treatment and control groups using a similar estimating equation as 5.4.1. As discussed earlier in this chapter, these can be especially useful for opening up the 'black box' of treatment effects. Illustrative examples of the value of such analysis include: Muralidharan and Sundararaman (2011) for the analysis of changes in teacher behavior in response to a teacher performance pay program; Das et al. (2013) for the analysis of changes in household spending in response to a school grant program; and Muralidharan and Sundararaman (2015) for the analysis of ways in which school time use differs markedly between public and private schools.

5.4.4 Cost Effectiveness

As described in section 2.3, a unifying theme in the economics of education literature is to compare the relative cost effectiveness of several possible interventions to improve education outcomes. Thus, the final piece of analysis that is recommended is a cost-effectiveness analysis that uses standardized templates for reporting cost (such as recommended by Dhaliwal et al. 2013) and presents treatment effects in terms of dollars spent per unit test score gain per student. An alternate form of "cost" effectiveness analysis that is not done often, but is also very useful is analyzing the effectiveness of interventions per unit of student time spent (see Muralidharan, Singh, and Ganimian 2016 for an illustration). While spending on education can (in theory) be augmented continuously, time is finite. Thus, identifying the effectiveness of interventions per unit of time spent is likely to play an important role in improving the productivity of education systems in developing countries.

6. Conclusion

The study of education in developing countries has benefited enormously from the rapid growth in field experiments. The most important learning from the experimental research in this area over the past fifteen years has been that there is a wide range of cost-effectiveness of education interventions. On the one hand, very expensive policies such as unconditional teacher salary increases have been found to have no impact on learning outcomes. On the other hand, relatively inexpensive policies like supplemental teaching at the right level with modestly paid and trained volunteers have been found to have large positive impacts on learning outcomes. Overall, the evidence points to several promising ways in which the efficiency of education spending in developing countries can be improved by pivoting public expenditure from less cost-effective categories of expenditure to more cost-effective ways of achieving the same objectives.

At the same time, there are important gaps in our knowledge of how best to design interventions to cost-effectively improve outcomes at scale, and there is much fertile ground for research on education in developing countries. There are important open questions within each of the four categories of interventions summarized in section 3. On demand, a lot more work is needed on the optimal design of demand-side interventions (beyond showing that they are effective). On inputs, we still do not have good experimental evidence on many important questions including the impacts of improving school infrastructure, teacher training programs, and class size (the evidence on this is more indirect). On governance, key open questions include understanding the impact of private schools (holding per-student spending constant, and precluding selection of students), and the impact of attempts to improve school governance at scale. In addition to these, I highlight three areas below where the knowledge gaps are large, and where the returns to better evidence are likely to be particularly high.

The first under-researched area where field experiments are likely to yield large returns is pedagogy. Most of the experimental studies on education in developing countries have been conducted by economists, and as a result the topics on which we have more evidence tend to be topics of interest to economists (such as household demand, information, inputs, and incentives). However, some of the most promising avenues for improving education in low-income settings may involve improving the design and delivery of classroom instruction. While several small scale innovations may be taking place in this area, there is remarkably little good evidence on the effectiveness of different pedagogical practices in developing countries. For instance, we have very little evidence on how to optimally organize a period of classroom instruction. In other words, the core building block of modern schooling (a period of instruction) is based on a non-experimental evidentiary standard. This in turn means that the evidence base for the design of teacher training programs is also very limited. While economists may have limited professional incentives to work on domains such as pedagogy, there are likely to be large social returns from researchers trained in designing and running field experiments collaborating with experts in curriculum and pedagogy to improve the empirical evidence base on effective pedagogy in developing countries.[37]

[37] A good illustration of this point is the consistent evidence on the large gains in student learning obtained from implementing a pedagogical approach that is focused on “Teaching at the Right Level”.

A second under-researched area is post-primary and secondary education. Most of the evidence summarized in this chapter has been from interventions aimed at improving primary education (with the notable exception of conditional cash transfer programs). This is understandable since primary education is foundational and most of the increases in developing country school enrollment in the past fifteen years have been in primary school (consistent with the MDG of achieving universal primary education by 2015). However, the cohorts who benefited from this expansion in primary education are now entering post-primary education in large numbers and there is remarkably little evidence on how to effectively improve the quality of post-primary education. The challenges of effective post-primary education are likely to be considerably greater than those of primary education since the variation in student preparation is likely to be much higher (as shown by Muralidharan, Singh, and Ganimian 2016), and the returns to research here are likely to be high.

A three under-researched area is scale. Specifically, while the evidence summarized in this paper provides a very useful starting point in identifying the kinds of interventions that are likely to be effective, we have very little understanding on how to embed these interventions in education systems to deliver improved outcomes at scale. One approach to developing such an understanding is provided by the research program on "Teaching at the Right Level" (TaRL) led by Abhijit Banerjee and Esther Duflo and conducted in partnership with Rukmini Banerjee of the NGO Pratham for over a decade. Starting with an "efficacy trial" on a small scale (reported in Banerjee et al. 2007), this research program has featured experimental evaluations of several implementation models in and outside the formal public school system across several Indian states to better understand how to improve basic literacy and numeracy at scale (see Banerjee et al. 2010, 2016 and Duflo et al. 2015 for a detailed discussion). While the studies reveal several challenges in successfully integrating the successful TaRL pedagogy into the regular education system, the long-term program of iterative program design, implementation, evaluation, and refinement, has helped to identify models (such as the learning camps in Uttar Pradesh, and the dedicated TaRL hour in Haryana) that may enable the scaling up of the successful TaRL intervention (see Banerjee et al. 2016 for a detailed discussion).

A second approach is to work with governments to randomize programs at scale to directly evaluate the impact of education programs at scale. This is the approach I take in ongoing work in the Indian state of Madhya Pradesh where the government agreed to randomize an ambitious program to improve school quality at the scale of 2,000 schools. Such an approach may be especially promising when combined with administrative data on outcomes which sharply reduces the cost of carrying out experiments at scale, and also makes it easier to conduct longer-term follow ups (see Hastings, Nielson, and Zimmermann 2015 for an illustration of such a study in Chile).

In addition to the topics mentioned above, researchers would also do well to pay attention to three cross-cutting themes across all categories of education interventions. These are (a) heterogeneity, (b) data on intermediate variables, and (c) combining experimental and structural methods. A recurring insight in the evidence to date is that optimal policies and interventions are likely to vary as a function of students' initial conditions, and interventions that cater effectively to such heterogeneity are likely to be more effective. A second recurring theme in the discussion in this chapter is the importance of collecting good data on intermediate variables to provide better insights on the mechanisms for program impact (or reasons for lack thereof). A third theme that younger researchers will benefit from paying attention to is the value of embedding experimental interventions in more general models of household behavior. Unifying the "treatment effects" approach outlined in section 2.3 and the more structural modeling approach outlined in section 2.4 is not easy, but if done well, such studies have the potential to expand the research and policy use of experiments, and provide a more informed basis for using experimental results to make predictions regarding the impact of variants of the policy studied.

Field experiments in education in developing countries have been an extremely fertile area of research and policy-relevant insights in the past fifteen years. This chapter has aimed to synthesize the most important insights from the existing research and to provide a toolkit for younger researchers embarking on answering the open questions in this area. I expect that the field will continue to be very active, and that it will produce several high quality studies in the years to come.

References:

Acemoglu, D. (2010). Theory, general equilibrium, and political economy in development economics. *The Journal of Economic Perspectives, 24*(3), 17-32.

Altonji, J. G., & Mansfield, R. K. (2014). Group-average observables as controls for sorting on unobservables when estimating group treatment effects: the case of school and neighborhood effects *(NBER Working Paper No. 20781)*. Cambridge, MA: National Bureau of Economic Research (NBER).

Andrabi, T., Das, J., & Khwaja, A. I. (2015). *Report cards: The impact of providing school and child test scores on educational markets*. (Policy Research Working Paper No. 7226). The World Bank, Washington, DC.

Andrabi, T., Das, J., Khwaja, A. I., & Zajonc, T. (2011). Do value-added estimates add value? Accounting for learning dynamics. *American Economic Journal: Applied Economics, 3*(3), 29-54.

Angelucci, M., & De Giorgi, G. (2009). Indirect effects of an aid program: How do cash transfers affect ineligibles' consumption? *The American Economic Review, 99*(1), 486-508.

Angrist, J., Bettinger, E., Bloom, E., King, E., & Kremer, M. (2002). Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment. *American Economic Review*, 1535-1558. doi:10.3386/w8343

Angrist, J., Bettinger, E., & Kremer, M. (2006). Long-term educational consequences of secondary school vouchers: Evidence from administrative records in Colombia. *American Economic Review, 96*, 847-862. doi:10.1257/aer.96.3.847

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (forthcoming). Teacher Quality and Learning Outcomes in Kindergarten. *The Quarterly Journal of Economics*.

Atkin, D. (2012). *Endogenous skill acquisition and export manufacturing in Mexico*. (NBER Working Paper No. 18266). National Bureau of Economic Research (NBER), Cambridge, MA.

Attanasio, O. P. (2015). *The determinants of human capital formation during the early years of life: Theory, measurement and policies*. European Economic Association (EEA), Toulouse, France.

Attanasio, O. P., Meghir, C., & Santiago, A. (2012). Education choices in Mexico: Using a structural model and a randomized experiment to evaluate Progresa. *The Review of Economic Studies, 79*(1), 37-66. doi:10.1093/restud/rdr015

Baird, S., Hicks, J. H., Kremer, M., & Miguel, E. (forthcoming). Worms at work: Long-run impacts of a child health investment. *The Quarterly Journal of Economics*.

Baird, S., McIntosh, C., & Ozler, B. (2011). Cash or condition? Evidence from a cash transfer experiment. *The Quarterly Journal of Economics, 126*, 1709-1753. doi:10.1093/qje/qjr032

Baker, G. P. (1992). Incentive contracts and performance measurement. *Journal of Political Economy*, 598-614.

Banerjee, A. V., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., Walton, M. (2016). *Mainstreaming an effective intervention: Evidence from randomized evaluations of "Teaching at the Right Level" in India*. Unpublished manuscript. Massachusetts Institute of Technology (MIT), Cambridge, MA.

Banerjee, A. V., Banerji, R., Duflo, E., Glennerster, R., & Khemani, S. (2010). Pitfalls of participatory programs: Evidence from a randomized evaluation in education in India. *American Economic Journal: Economic Policy, 2*, 1-30. doi:10.1257/pol.2.1.1

Banerjee, A. V., Banerji, R., Duflo, E., & Walton, M. (2011). What helps children to learn? Evaluation of Pratham's Read India program in Bihar & Uttarakhand. Cambridge, MA: Abdul Latif Jameel Poverty Action Lab (JPAL).

Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. *The Quarterly Journal of Economics, 122*, 1235-1264. doi:10.1162/qjec.122.3.1235

Barlevy, G., & Neal, D. (2012). Pay for percentile. *American Economic Review, 102*(5), 1805-1831.

Barrera-Osorio, F., Bertrand, M., Linden, L. L., & Perez-Calle, F. (2011). Improving the design of conditional transfer programs: Evidence from a randomized education experiment in Colombia. *American Economic Journal: Applied Economics, 3*, 167-195. doi:10.1257/app.3.2.167

Barrera-Osorio, F., & Linden, L. (2009). *The use and misuse of computers in education: Evidence from a randomized experiment in Colombia*. (Impact Evaluation Series No. 29). The World Bank, Washington, DC.

Barrera-Osorio, F., Linden, L. L., & Saavedra, J. E. (2016). *Medium term educational consequences of alternative conditional cash transfer designs: Experimental evidence from Colombia*. Harvard Graduate School of Education, Cambridge, MA.

Barro, R. J. (1991). Economic growth in a cross section of countries. *The Quarterly Journal of Economics, CVI*(425), 407-443.

Beasley, E., & Huillery, E. (2012). *Empowering Parents in Schools: What They Can(not) Do*. Unpublished manuscript. Abdul Latif Jameel Poverty Action Lab (J-PAL), Cambridge, MA.

Becker, G. S. (1962). Investment in human capital: A theoretical analysis. *The journal of political economy*, 9-49.

Ben-Porath, Y. (1967). The production of human capital and the life cycle of earnings. *The journal of political economy*, 352-365.

Benhassine, N., Devoto, F., Duflo, E., Dupas, P., & Pouliquen, V. (2013). Turning a shove into a nudge? A “labeled cash transfer” for education. *American Economic Journal: Economic Policy, 7*, 86-125. doi:10.1257/pol.20130225

Berry, J. (2015). Child control in education decisions: An evaluation of targeted incentives to learn in India. *The Journal of Human Resources, 50*(4), 1051-1080.

Beuermann, D. W., Cristia, J. P., Cruz-Aguayo, Y., Cueto, S., & Malamud, O. (2015). Home computers and child outcomes: Short-term impacts from a randomized experiment in Peru. *American Economic Journal: Applied Economics, 7*(2), 53-80.

Bharadwaj, P., Løken, K. V., & Neilson, C. (2013). Early life health interventions and academic achievement. *The American Economic Review, 103*(5), 1862-1891.

Blattman, C., & Niehaus, P. (2014). Show Them the Money: Why Giving Cash Helps Alleviate Poverty. *Foreign Aff., 93*, 117.

Bleakley, H. (2007). Disease and development: evidence from hookworm eradication in the American South. *The Quarterly Journal of Economics, 122*(1), 73.

Blimpo, M. P. (2014). Team incentives for education in developing countries: A randomized field experiment in Benin. *American Economic Journal: Applied Economics, 6*(4), 90-109. doi:10.1257/app.6.4.90

Bobba, M., & Frisancho, V. (2016). *Learning about oneself: The effects of signaling academic ability on school choice*. Unpublished manuscripts. Inter-American Development Bank, Washington, DC.

Bobonis, G. J. (2009). Is the allocation of resources within the household efficient? New evidence from a randomized experiment. *Journal of Political Economy, 117*(3), 453-503.

Bobonis, G. J., & Finan, F. (2009). Neighborhood peer effects in secondary school enrollment decisions. *The Review of Economics and Statistics, 91*(4), 695-716.

Bold, T., Kimenyi, M., Mwabu, G., Ng'ang'a, A., & Sandefur, J. (2013). *Scaling-up what works: Experimental evidence on external validity in Kenyan education*. Unpublished manuscript. Center for Global Development, Washington, DC.

Borkum, E., He, F., & Linden, L. L. (2013). *School libraries and language skills in Indian primary schools: A randomized evaluation of the Akshara Library program*. Unpublished manuscript. Abdul Latif Jameel Poverty Action Lab (J-PAL), Cambridge, MA.

Bourdon, J., Frölich, M., & Michaelowa, K. (2010). Teacher shortages, teacher contracts and their effect on education in Africa. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 173*(1), 93-116.

Bruhn, M., & McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American Economic Journal: Applied Economics, 1*(4), 200-232.

Bruns, B., & Luque, J. (2014). *Great teachers: How to raise student learning in Latin America and the Caribbean*. Washington, DC: The World Bank.

Burde, D., & Linden, L. L. (2013). The effect of village-based schools: Evidence from a randomized controlled trial in Afghanistan. *American Economic Journal: Applied Economics, 5*, 27–40. doi:10.1257/app.5.3.27

Carneiro, P., Heckman, J. J., & Vytlacil, E. J. (2011). Estimating marginal returns to education. *The American Economic Review, 101*(6), 2754-2781.

Cartwright, N. (2007). Are RCTs the gold standard? *BioSocieties, 2*(1), 11-20.

Chaudhury, N., Hammer, J., Kremer, M., Muralidharan, K., & Rogers, F. H. (2006). Missing in action: Teacher and health worker absence in developing countries. *The Journal of Economic Perspectives, 20*(1), 91-116.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from project STAR. *The Quarterly Journal of Economics, 126*, 1593-1660. doi:10.1093/qje/qjr041

Conn, K. (2014). *Identifying effective education interventions in Sub-Saharan Africa: A meta-analysis of rigorous impact evaluations*. Unpublished manuscript. Columbia University, New York, NY.

Cristia, J., Ibarrarán, P., Cueto, S., Santiago, A., & Severín, E. (2012). *Technology and child development: Evidence from the One Laptop per Child program*. (Working Paper No. IDB-WP-304). Inter-American Development Bank, Washington, DC.

Das, J., Dercon, S., Habyarimana, J., Krishnan, P., Muralidharan, K., & Sundararaman, V. (2013). School inputs, household substitution, and test scores. *American Economic Journal: Applied Economics, 5*, 29-57. doi:10.1257/app.5.2.29

de Ree, J., Muralidharan, K., Pradhan, M., & Rogers, F. H. (2016). *Double for nothing? Experimental evidence on the impact of an unconditional teacher salary increase on student performance in Indonesia*. Unpublished manuscript. The World Bank, Washington, DC.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature, 48*(2), 424-455.

Dehejia, R., Pop-Eleches, C., & Samii, C. (2015). *From local to global: External validity in a fertility natural experiment*. Wagner Graduate School of Public Service, New York, NY.

Dhaliwal, I., Duflo, E., Glennerster, R., & Tulloch, C. (2012). *Comparative cost-effectiveness analysis to inform policy in developing countries: A general framework with applications for education*. Unpublished manuscript. Abdul Latif Jameel Poverty Action Lab (J-PAL), Cambridge, MA.

Dizon-Ross, R. (2016). *Parents' perceptions and children's education: Experimental evidence from Malawi*. Unpublished manuscript. Massachusetts Institute of Technology (MIT), Cambridge, MA.

Duflo, E. (2001). Schooling and labor market consequences of school construction in Indonesia: Evidence from an unusual policy experiment. *American Economic Review, 91*, 795-813. doi:10.1257/aer.91.4.795

Duflo, E., Berry, J., Mukerji, S., & Shotland, M. (2015a). *A wide angle view of learning: Evaluation of the CCE and LEP programmes in Haryana, India*. (Impact Evaluation Report No. 22). International Initiative for Impact Evaluation (3ie), New Delhi, India.

Duflo, E., Dupas, P., & Kremer, M. (2011). Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. *The American Economic Review, 101*, 1739-1774. doi:10.1257/aer.101.5.1739

Duflo, E., Dupas, P., & Kremer, M. (2015b). School governance, teacher incentives, and pupil-teacher ratios: Experimental evidence from Kenyan primary schools. *Journal of Public Economics, 123*, 92–110. doi:10.1016/j.jpubeco.2014.11.008

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of development economics, 4*, 3895-3962.

Duflo, E., Hanna, R., & Ryan, S. P. (2012). Incentives work: Getting teachers to come to school. *The American Economic Review, 102*, 1241-1278. doi:10.1257/aer.102.4.1241

Evans, D. K., & Popova, A. (2015). *What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews*. (Policy Research Working Paper No. 7203). The World Bank, Washington, DC.

Fiszbein, A., & Schady, N. R. (2009). *Conditional Cash Transfers: Reducing Present and Future Poverty*. Washington, DC: The World Bank.

Foster, A. D., & Rosenzweig, M. R. (1995). Learning by doing and learning from others: Human capital and technical change in agriculture. *Journal of Political Economy*, 1176-1209.

Ganimian, A. J., & Murnane, R. J. (2016). Improving educational outcomes in developing countries: Lessons from rigorous evaluations. *Review of Educational Research, XX*(X), 1-37.

Gerber, A. S., & Green, D. P. (2012). *Field Experiments: Design, Analysis, and Interpretation*. New York, NY: W.W. Norton.

Gertler, P., Heckman, J., Pinto, R., Zanolini, A., Vermeerch, C., Walker, S., . . . Grantham-McGregor, S. (2014). Labor market returns to an early childhood stimulation intervention in Jamaica. *Science, 344*(6187), 998-1001.

Glennerster, R., & Takavarasha, K. (2013). *Running randomized evaluations: A practical guide*: Princeton University Press.

Glewwe, P., Hanushek, E. A., Humpage, S. D., & Ravina, R. (2014). School resources and educational outcomes in developing countries: A review of the literature from 1990 to 2010. In P. Glewwe (Ed.), *Education Policy in Developing Countries*. Chicago, IL and London, UK: University of Chicago Press.

Glewwe, P., Ilias, N., & Kremer, M. (2010). Teacher incentives. *American Economic Journal: Applied Economics, 2*, 205-227. doi:10.1257/app.2.3.205

Glewwe, P., Kremer, M., & Moulin, S. (2009). Many children left behind? Textbooks and test scores in Kenya. *American Economic Journal: Applied Economics, 1*, 112-135. doi:10.1257/app.1.1.112

Glewwe, P., Kremer, M., Moulin, S., & Zitzewitz, E. (2004). Retrospective vs. prospective analyses of school inputs: The case of flip charts in Kenya. *Journal of Development Economics, 74*, 251-268. doi:10.1016/j.jdeveco.2003.12.010

Glewwe, P., & Maïga, E. W. (2011). The impacts of school management reforms in Madagascar: Do the impacts vary by teacher type? *Journal of development effectiveness, 3*(4), 435-469. doi:10.1080/19439342.2011.604729

Glewwe, P., & Muralidharan, K. (2016). *Improving school education outcomes in developing countries: Evidence, knowledge gaps, and policy implications*. Handbook of economics of education.

Green, J. R., & Stokey, N. L. (1983). A comparison of tournaments and contracts. *The journal of political economy*, 349-364.

Hastings, J., Neilson, C. A., Zimmerman, S. D. (2015). The effect of earnings disclosure on college enrollment decisions. (Working Paper No. 21300). National Bureau of Economic Research (NBER), Cambridge, MA.

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, A. (2010). The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics, 94*(1), 114-128.

Heckman, J. J., & Smith, J. A. (1995). Assessing the case for social experiments. *The Journal of Economic Perspectives, 9*(2), 85-110.

Hirshleifer, S. (2015). Incentives for effort or outputs? A field experiment to improve student performance *Unpublished manuscript*. San Diego, CA: University of California at San Diego.

Holmstrom, B., & Milgrom, P. (1987). Aggregation and linearity in the provision of intertemporal incentives. *Econometrica: Journal of the Econometric Society*, 303-328.

Holmstrom, B., & Milgrom, P. (1991). Multitask principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, & Organization, 7*, 24-52.

Imbens, G., & Athey, S. (2016). The econometrics of randomized experiments. In E. Duflo & A. Banerjee (Eds.), *Handbook of field experiments*: North Holland.

Itoh, H. (1991). Incentives to help in multi-agent situations. *Econometrica: Journal of the Econometric Society*, 611-636.

Jensen, R. (2010). The (perceived) returns to education and the demand for schooling. *The Quarterly Journal of Economics, 125*, 515-548. doi:10.1162/qjec.2010.125.2.515

Jensen, R. (2012). Do labor market opportunities affect young women's work and family decisions? Experimental evidence from India. *The Quarterly Journal of Economics, 127*, 753–792. doi:10.1093/qje/qjs002

Kandel, E., & Lazear, E. P. (1992). Peer pressure and partnerships. *Journal of Political Economy*, 801-817.

Kandori, M. (1992). Social norms and community enforcement. *The Review of Economic Studies, 59*(1), 63-80.

Kling, J., Ludwig, J., Congdon, B., & Mullainathan, S. (2016). Social policy: Mechanism experiments and policy evaluations. In E. Duflo & A. Banerjee (Eds.), *Handbook of field experiments*: North Holland.

Kremer, M. (1993). The O-ring theory of economic development. *The Quarterly Journal of Economics*, 551-575.

Kremer, M. (2003). Randomized evaluations of educational programs in developing countries: Some lessons. *The American Economic Review, 93*(2), 102-106.
Kremer, M., Brannen, C., & Glennerster, R. (2013). The challenge of education and learning in the developing world. *Science, 340*, 297-300. doi:10.1126/science.1235350
Kremer, M., Miguel, E., & Thornton, R. (2009). Incentives to learn. *The Review of Economics and Statistics, 91*, 437-456. doi:10.1162/rest.91.3.437
Kremer, M., & Sarychev, A. (2008). *Why do governments operate schools?* Unpublished manuscript. Harvard University, Cambridge, MA.
Krishnaratne, S., White, H., & Carpenter, E. (2013). *Quality education for all children? What works in education in developing countries*. (Working Paper No. 20). International Initiative for Impact Evaluation (3ie), New Delhi, India.
Lai, F., Luo, R., Zhang, L., Huang, X., & Rozelle, S. (2015). Does computer-assisted learning improve learning outcomes? Evidence from a randomized experiment in migrant schools in Beijing. *Economics of Education Review, 47*, 34-48. doi:10.1016/j.econedurev.2015.03.005
Lakshminarayana, R., Eble, A., Bhakta, P., Frost, C., Boone, P., Elbourne, D., & Mann, V. (2013). The Support to Rural India's Public Education System (STRIPES) trial: A cluster randomised controlled trial of supplementary teaching, learning material and material support. *PloS one, 8*(7), e65775.
Lalive, R., & Cattaneo, M. A. (2009). Social interactions and schooling decisions. *The Review of Economics and Statistics, 91*(3), 457-477.
LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 604-620.
Lassibille, G., Tan, J.-P., Jesse, C., & Nguyen, T. V. (2010). Managing for results in primary education in Madagascar: Evaluating the impact of selected workflow interventions. *The World Bank Economic Review*, 1-27. doi:10.1093/wber/lhq009
Lazear, E. P. (2006). Speeding, terrorism, and teaching to the test. *The Quarterly Journal of Economics*, 1029-1061.
Lazear, E. P., & Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *Journal of Political Economy, 89*(5), 841-864.
Linden, L. L. (2008). *Complement or substitute? The effect of technology on student achievement in India*. Unpublished manuscript. Abdul Latif Jameel Poverty Action Lab (J-PAL), Cambridge, MA.
Loyalka, P., Liu, C., Song, Y., Yi, H., Huang, X., Wei, J., . . . Rozelle, S. (2013). Can information and counseling help students from poor rural areas go to high school? Evidence from China. *Journal of Comparative Economics, 41*, 1012–1025. doi:10.1016/j.jce.2013.06.004
Lucas, R. E. (1988). On the mechanics of economic development. *Journal of Monetary Economics, 22*(1), 3-42.
Lucas, R. E. (1990). Why doesn't capital flow from rich to poor countries? *The American Economic Review, 80*(2), 92-96.
Malamud, O., & Pop-Eleches, C. (2011). Home computer use and the development of human capital. *The Quarterly Journal of Economics, 126*, 987-1027. doi:10.1093/qje/qjr008
Mankiw, N. G. (2006). The macroeconomist as scientist and engineer. *The Journal of Economic Perspectives, 20*(4), 29-46.

Mankiw, N. G., Romer, D., & Weil, D. N. (1992). A contribution to the empirics of economic growth. *The Quarterly Journal of Economics, 107*(2), 407-437.

Mbiti, I., Muralidharan, K., Romero, M., Schipper, Y., Manda, C., & Rajani, R. (2016). *Inputs, incentives, and complementarities in primary education: Experimental evidence from Tanzania*. Unpublished manuscript. University of California at San Diego, San Diego, CA.

McEwan, P. (2014). Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments. *Review of Educational Research, XX*, 1-42. doi:10.3102/0034654314553127

Miguel, E., & Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica, 72*, 159-217. doi:10.1111/j.1468-0262.2004.00481.x

Moretti, E. (2004). Workers' education, spillovers, and productivity: evidence from plant-level production functions. *The American Economic Review, 94*(3), 656-690.

Muralidharan, K. (2012). *Long-term effects of teacher performance pay: Experimental evidence from India*. Unpublished manuscript. University of California, San Diego, San Diego, CA.

Muralidharan, K. (2013). Priorities for primary education policy in India's 12th five-year plan. *India Policy Forum 2012-13, 9*, 1-46.

Muralidharan, K. (2016). *A new approach to public sector hiring in India for improved service delivery*. University of California at San Diego, San Diego, CA.

Muralidharan, K., Das, J., Holla, A., & Mohpal, A. (2016a). *The fiscal cost of weak governance: Evidence from teacher absence in India*. Unpublished manuscript. University of California, San Diego, San Diego, CA.

Muralidharan, K., Niehaus, P., & Sukhtankar, S. (forthcoming). Building state capacity: Evidence from biometric smartcards in India. *American Economic Review*.

Muralidharan, K., & Prakash, N. (2016). *Cycling to school: Increasing secondary school enrollment for girls in India*. Unpublished manuscript. University of California at San Diego, San Diego, CA.

Muralidharan, K., Singh, A., & Ganimian, A. J. (2016b). *Teaching all students, and not just the top of the class: Experimental evidence on technology-led education in India*. Unpublished manuscript. University of California at San Diego, San Diego, CA.

Muralidharan, K., & Sundararaman, V. (2010). The impact of diagnostic feedback to teachers on student learning: Experimental evidence from India. *The Economic Journal, 120*, F187-F203. doi:10.1111/j.1468-0297.2010.02373.x

Muralidharan, K., & Sundararaman, V. (2011). Teacher performance pay: Experimental evidence from India. *The journal of political economy, 119*, 39-77. doi:10.1086/659655

Muralidharan, K., & Sundararaman, V. (2013). *Contract teachers: Experimental evidence from India*. (Working Paper No. 19440). National Bureau of Economic Research (NBER), Cambridge, MA.

Muralidharan, K., & Sundararaman, V. (2015). The aggregate effect of school choice: Evidence from a two-stage experiment in India. *The Quarterly Journal of Economics, 130*(3), 1011-1066. doi:10.1093/qje/qjv013

Muralidharan, K., & Zieleniak, Y. (2014). *Chasing the syllabus: Measuring learning trajectories in developing countries with longitudinal data and item response theory*. Unpublished manuscript. University of California, San Diego, San Diego, CA.

Neal, D., & Schanzenbach, D. W. (2010). Left behind by design: Proficiency counts and test-based accountability. *The Review of Economics and Statistics, 92*(2), 263-283.
Olken, B. A. (2015). Promises and perils of pre-analysis plans. *The Journal of Economic Perspectives, 29*(3), 61-80.
Pradhan, M., Suryadarma, D., Beatty, A., Wong, M., Gaduh, A., Alisjahbana, A., & Artha, R. P. (2014). Improving educational quality through enhancing community participation: Results from a randomized field experiment in Indonesia. *American Economic Journal: Applied Economics, 6*, 105-126. doi:10.1257/app.6.2.105
Pritchett, L., & Sandefur, J. (2013). *Context matters for size: Why external validity claims and development practice don't mix*. (Working Paper No. 336). Center for Global Development (CGD), Washington, DC.
Rawal, S., & Kingdon, G. (2010). Akin to my teacher: Does caste, religious or gender distance between student and teacher matter? Some evidence from India: Department of Quantitative Social Science-UCL Institute of Education, University College London.
Reinikka, R., & Svensson, J. (2011). The power of information in public services: Evidence from education in Uganda. *Journal of Public Economics, 95*(7), 956-966. doi:10.1016/j.jpubeco.2011.02.006
Sabarwal, S., Evans, D., & Marshak, A. (2013). *The permanent textbook hypothesis: School inputs and student outcomes in Sierra Leone*. (Policy Research Working Paper No. 7021). The World Bank, Washington, DC.
Sen, A. (1993). Capability and well being. In A. Sen & M. Nussbaum (Eds.), *The quality of life*. Oxford, United Kingdom: Oxford University Press.
Singh, A. (2015). Private school effects in urban and rural India: Panel estimates at primary and secondary school ages *Journal of Development Economics, 113*, 16-32.
Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal, 113*(485), F3-F33.
Todd, P. E., & Wolpin, K. I. (2006). Assessing the impact of a school subsidy program in Mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility. *The American Economic Review, 96*(5), 1384-1417.
Yang, Y., Zhang, L., Zeng, J., Pang, X., Lai, F., & Rozelle, S. (2013). Computers and the academic performance of elementary school-aged girls in China's poor communities. *Computers & Education, 60*(1), 335-346. doi:10.1016/j.compedu.2012.08.011